Integral comparison of static and dynamic ovarian reserve tests; a prospective study and a systematic review

Kwee, J.

2007

document version
Publisher's PDF, also known as Version of record

Integral comparison of static and dynamic ovarian reserve tests; a prospective study and a systematic review

Kwee, Janet
Integral comparison of static and dynamic ovarian reserve tests; a prospective study and a systematic review
Thesis Vrije Universiteit Amsterdam. With summary in Dutch.

The research presented in this thesis was performed at the IVF Centre, Division of Reproductive Medicine, Department of Obstetrics and Gynaecology, Vrije Universiteit medical centre, Amsterdam, The Netherlands.

ISBN: 90-8659-0691

Cover: The principles of the Exogenous Ovarian Reserve Test, designed by Nils Lambalk
Printed by Gildeprint Drukkerijen B.V., Enschede, The Netherlands

VRIJE UNIVERSITEIT

Integral comparison of static and dynamic ovarian reserve tests; a prospective study and a systematic review

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op vrijdag 15 februari 2007 om 10.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Janet Kwee

geboren te Winschoten

promotor: prof.dr. R. Homburg
copromotor: dr. C.B. Lambalk

De wereld is voor iedereen
Tjin Hwa Kwee (vrij naar Confucius)

Aan mijn ouders
Aan Paul

Contents

Chapter 1	Introduction and aim of the thesis	xx
Chapter 2	Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment : results of a prospective randomized study. *Human Reproduction 2003;18:1422-7*	xx
Chapter 3	The Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of low and hyperresponders to in vitro fertilization (IVF). *Fertility Sterility, 2006;85:1714-22*	xx
Chapter 4	Intercycle variability of ovarian reserve tests: results of a prospective randomized study. *Human Reproduction 2004;19:590-5*	xx
	Reply: Variability of ovarian reserve tests *Human Reproduction 2004;19:2170-1*	xx
Chapter 5	Ovarian volume and antral follicle count for the prediction of low and hyper responders with in vitro fertilization. *Submitted*	xx
Chapter 6	Evaluation of AMH as a test for the prediction of ovarian reserve. *Submitted*	xx
Chapter 7	A systematic review of tests predicting ovarian reserve and IVF outcome. *Human Reproduction Update 2006;6:685-718*	xx
Chapter 8	General discussion, conclusions and recommendations for future research	xx
	Summary	xx
	Summary in Dutch	xx
	Acknowledgements	xx
	List of publications	xx
	Curriculum vitae	xx

Chapter 1

Introduction

Introduction

The percentage of the general population seeking help for infertility is growing. One of the reasons for this is the fact that, especially in the western world, women are postponing their pregnancies because of their career. Women, starting with an infertility work-up, undergo extensive testing, and a large proportion of them will require expensive and invasive therapies, including assisted reproductive technologies. This introduction reviews static and dynamic ovarian function tests, which supposedly can predict ovarian reserve. Ovarian reserve is currently defined as the number and quality of the follicles left at any moment in the ovary. An accurate measure of the quantitative ovarian reserve would be the counting of all follicles present in both ovaries, as done in post mortem studies (Block, 1952). For obvious reasons, in ovarian reserve testing the true size of the follicle pool has not been used as the gold standard for evaluation (Lambalk *et al.*, 2004, Sharara and Scott, 2004, Lass *et al.*, 1997a, Lass, 2004), apart from one distinct study (Gulekli *et al.*, 1999), where whole ovary counts served as reference for several ovarian reserve tests. Instead, several proxy variables of the pool size are used in studies on diagnostic accuracy, like ovarian response to hyperstimulation with exogenous FSH in IVF, and the occurrence of menopause or menopausal transition, as these events are highly quantity determined. Although related, the quality of the oocyte released from the dominant follicle at ovulation represents the other aspect of ovarian reserve. Proxy variables for oocyte quality currently used are the pregnancy probability in infertility treatment like IUI and IVF or in the follow up of couples during and after the initial infertility work-up.
The purpose of this introduction is to examine the clinical tools currently available to assess ovarian reserve leading to a prognosis of the reproductive potential of a woman.

Predictors of ovarian reserve

Age

The ovarian reserve diminishes with increasing age. The value of age in predicting the outcome of assisted reproduction has been observed in several studies (Hughes *et al.*, 1989, Meldrum, 1993, Navot *et al.*, 1994, Scott *et al.*, 1995). The number of oocytes retrieved after ovarian hyperstimulation and pregnancy rates are inversely correlated with increasing age (Toner *et al.*, 1991). Because in donor oocyte programs, recipient's age is not correlated with pregnancy rates, fecundity is almost independent of uterine age, and, almost completely dependent on the donor's ovarian age and reserve (Meldrum, 1993, Navot *et al.*, 1994). However, the biological age of the ovaries does not always match a woman's calendar age. Decreasing ovarian reserve occurs at different ages. This illustrates that calendar age alone is not a sufficient parameter for judging the individual ovarian reserve.

Cycle length

During the perimenopausal period the cycle becomes irregular. The duration of this period of irregular cyclity, during which unusually long and short cycles are often interspersed, varies greatly among women (Treloar *et al.*, 1967). Analysis of the hormonal changes during the menopausal transition emphasizes the complexity of the hypothalamic-pituitary-gonadal regulatory system. The deficiency in the production of ovarian steroids and peptides leads to the fact that FSH levels are intermittently elevated in perimenopausal women (Sherman *et al.*,

1976). The elevated concentrations of FSH may be responsible for changes in the follicular phase including shorter length consistent with an earlier start of growth of the follicles. In contrast, luteal phase length and progesterone production essentially are unchanged until very late in the aging process (Sherman *et al.*, 1976).
Although FSH concentrations begin to rise significantly from age 40 onwards, follicular phase length does not shorten significantly until the age of 44 (Lenton *et al.*, 1984, Lenton *et al.*, 1988). There is a small but steady decline in follicular phase length throughout reproductive life. The lengthening which occurs from the age of 46 may indicate the first sign of ovarian insensitivity to the raised gonadotropins and the approach of the impending peri-menopausal period, in which the menstrual cycle length, though very variable, is often prolonged (Lenton *et al.*, 1988).
Theoretically there should be a correlation. In fact, practically, follicular cycle length has no better predictive value than the calendar age of a women.
The limited predictive value of age alone in assessing a given individual's chances for conception led to evaluation of the predictive value of other parameters. Recently, several investigators have described tests predicting the ovarian response to gonadotropins.
.

Ovarian Biopsy

Ovarian reserve depends on the number of primordial follicles in the ovarian cortex which suggests that the obvious way to obtain an estimate would be to measure follicular density in an ovarian biopsy (Lass *et al.*, 2001, Lass, 2004). Attempts were made to quantify the number of small antral follicles in small shallow biopsies taken during diagnostic laparoscopy from infertility patients (Lass *et al.*, 1997a) and there was a clear age dependent decline in follicular density. Women over 35 years of age had only 30 % of the quantities present in younger women. The number of follicles per unit of volume found in the biopsies was used to estimate the total and it was suggested that it could as such potentially be applied at the individual level. It was recognized though that biopsy for follicle density would not accurately represent the density in the whole ovary (Lass, 2001) and this seems indeed the case. Recently, several investigators have shown that follicle density varied greatly in small pieces of cortex rendering information from biopsies as completely unreliable for an individual ovarian follicle content irrespective of how many were taken, their size and the location (Lambalk *et al.*, 2004, Qu *et al.*, 2000, Schmidt *et al.*, 2003). This indicates that the technique which is invasive and potentially harmful in terms of risks for adhesions and other complications of the surgical procedure is intrinsically unreliable and should therefore not be used to evaluate individual ovarian reserve. It is probably useful for research purposes to determine follicle density statistics in patient groups provided that group sizes are such that they compensate for the inherent extreme interbiopsy and inter-individual spread of information (Lambalk *et al.*, 2004, Webber *et al.*, 2003, Qu *et al.*, 2000, Schmidt *et al.*, 2003).
Finally, in the context of the current systematic review there are no studies published that evaluated ovarian biopsy follicle density for prediction of IVF outcome in terms of ovarian response and pregnancy rates.

Basal FSH

Over the past decade, several markers of ovarian aging have been identified. The earliest and most consistent reproductive endocrine finding associated with reproductive aging in women is an isolated excessive rise in circulating FSH levels during the luteal-follicular transition (Sherman *et al.*, 1976). This rise in FSH in the normal menstrual cycle is essential for recruitment of a cohort of follicles into the developing pool from which a dominant follicle will ultimately be selected.

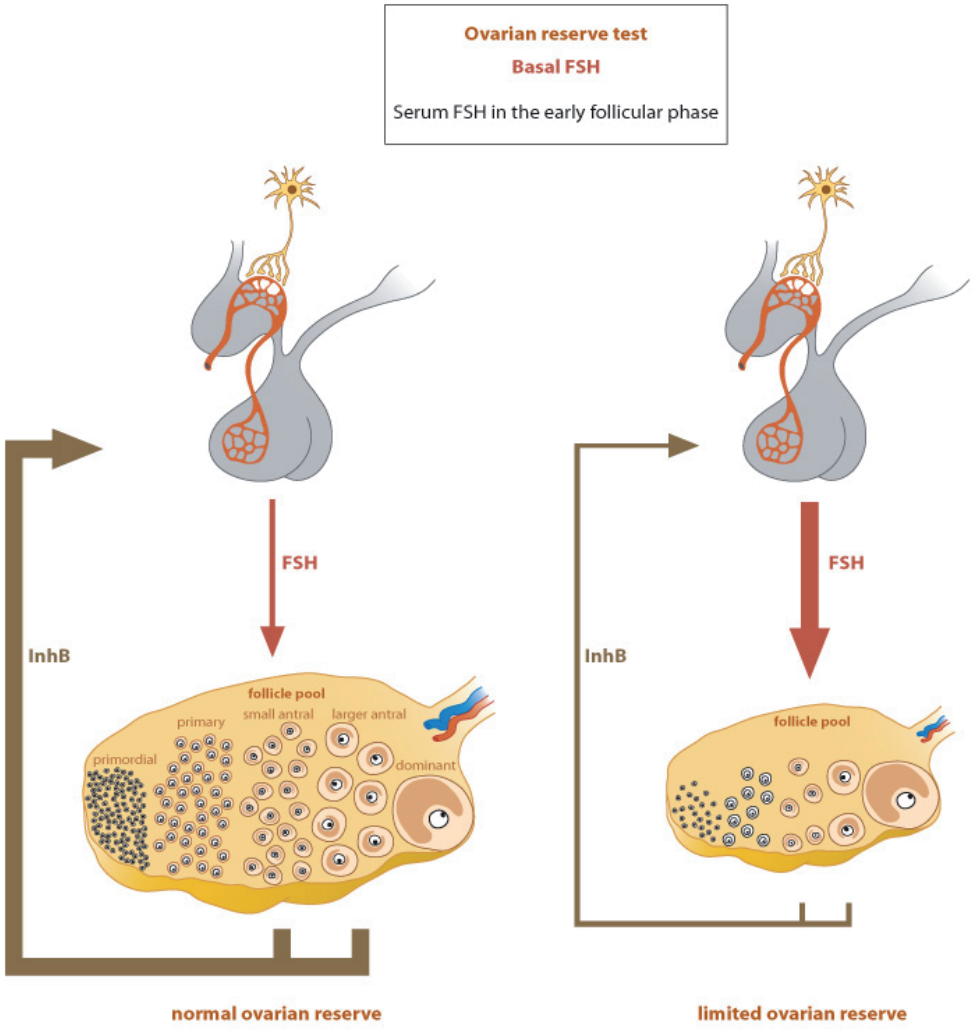
Several studies show (Pearlstone *et al.*, 1992, Toner, 1993, Cahill *et al.*, 1994, Hansen *et al.*, 1996) that the basal FSH level (bFSH), measured on cycle day 3, has a better predictive value than the calendar age of a woman. The factors responsible for the elevation in early follicular phase FSH levels during the perimenopausal phase are only partially known. However, it is well accepted that bFSH is an indirect measure of the ovarian reserve (Scott, 1989). It is a reflection of the balance between ovarian steroids and peptides on the one side and the hypothalamic GnRH stimulation on the other side during the period of follicular recruitment at the time just before the selection of the dominant follicle. It signifies the amount of inhibin and/or E2 produced by the cohort of follicles, responsible for the negative feedback on the FSH secretion. The bFSH increases when ovarian reserve diminishes (Lenton *et al.*, 1988), supposedly because the small antral follicles produce less inhibin B and possibly E2 (fig. 1).

Scott et al. (Scott *et al.*, 1990) evaluated the intercycle variability and its potential impact on the response to hyperstimulation with FSH. Women with a normal bFSH had a small range in the intercycle variability, in contrast to women with an elevated bFSH showing a much greater variation. The diagnostic and predictive value of a single determination of the bFSH is therefore limited, nor is it a good parameter for selecting the optimal cycle for ovarian hyperstimulation. If the patient has wide fluctuations in her basal FSH values, she is more likely to respond poorly to hyperstimulation (Scott *et al.*, 1990). This information should be useful in counseling patients with respect to chances in assisted reproduction. Martin et al. (1996) showed that if the bFSH at a single occasion was elevated, ovarian response to gonadotropins was poor in every cycle.

Khalifa *et al.* (1992) showed that in the presence of only one ovary the bFSH also has a diagnostic value, although the mean bFSH is higher in women with one ovary compared with those with two, respectively 17,3 ± 9,1 IU/L and 12,1 ± 3,3 IU/L (normal bFSH level ≤ 15 IU/L). Lambalk et al. (1998) showed in mothers of dizygotic twins that the mean bFSH levels are higher (9,8 ± 5,5 IU/L versus 6,0 ± 3,0 IU/L, normal bFSH level ≤ 10 IU/L) compared to singleton mothers. This phenomenon explains the occurrence of multifollicular growth leading to more than one ovulation per cycle resulting in a higher chance for a multiple pregnancy (Beemsterboer *et al.*, 2006). This phenomenon is genetically determined. The importance of these findings is that not all patients with an elevated bFSH have a diminished ovarian reserve.

Bancsi *et al.* (2003), examined 21 studies that reported on bFSH and IVF outcome. He concluded after a metanalysis that possible clinical value of bFSH is restricted to a small minority of patients i.e. bFSH should not be regarded as a useful routine test in the prediction of IVF outcome.

Figure 1. The principles of the ovarian reserve test: bFSH

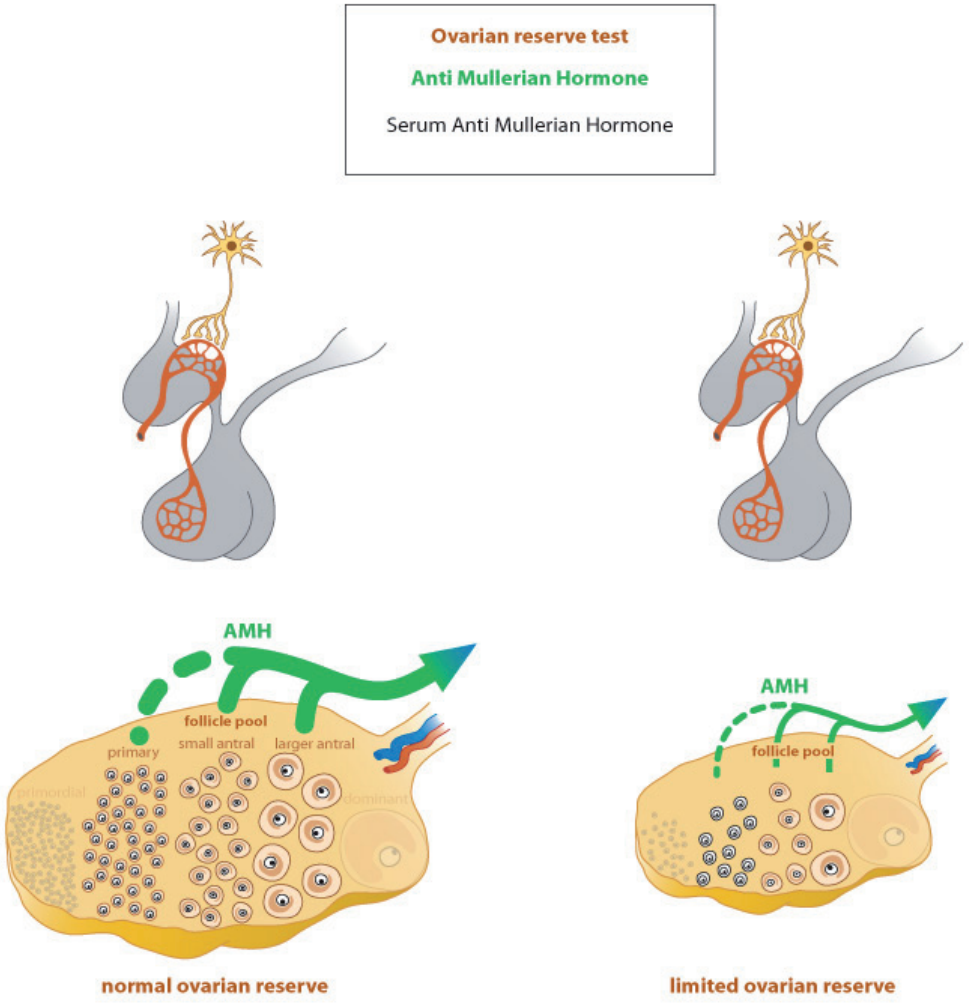


Anti-Müllerian hormone (AMH)

Anti-Müllerian hormone, also known as Müllerian Inhibiting Substance (MIS), is a dimeric glycoprotein belonging to the Transforming Growth Factors-β (TGF-β) family. It is involved in the regression of the Müllerian ducts during male fetal development (Behringer et al., 1994) and expressed in Sertoli cells from testicular differentation up to puberty. In females, AMH is exclusively produced by granulosa cells of preantral (primary and secondary) and small antral follicles (Vigier *et al.*, 1984) from birth up to menopause (fig. 2). After follicles differentiate from the primordial to the primary stage, production of AMH starts and it continues until the follicles have reached the antral stages with diameters of 2-6 mm (Durlinger *et al.*, 2001, Durlinger *et al.*, 1999). The number of the small antral follicles is related to the size of the primordial follicle pool (Gougeon, 1984). With the decrease in the

number of the antral follicles with age AMH production appears to become diminished (van Rooij et al., 2004, van Rooij et al., 2005, de Vet et al., 2002) and will invariably become undetectable at and after menopause.
Studies in IVF stimulations have suggested that AMH as such represents ovarian quantitative reserve (van Rooij et al., 2002, Seifer et al., 2002, Fanchin, et al., 2003). Moreover, evidence is accumulating that AMH, in contrast to FSH, Estradiol and InhibinB, can be used as a cycle independent marker (Hehenkamp et al., 2006).

Figure 2. The principles of the ovarian reserve test: AMH

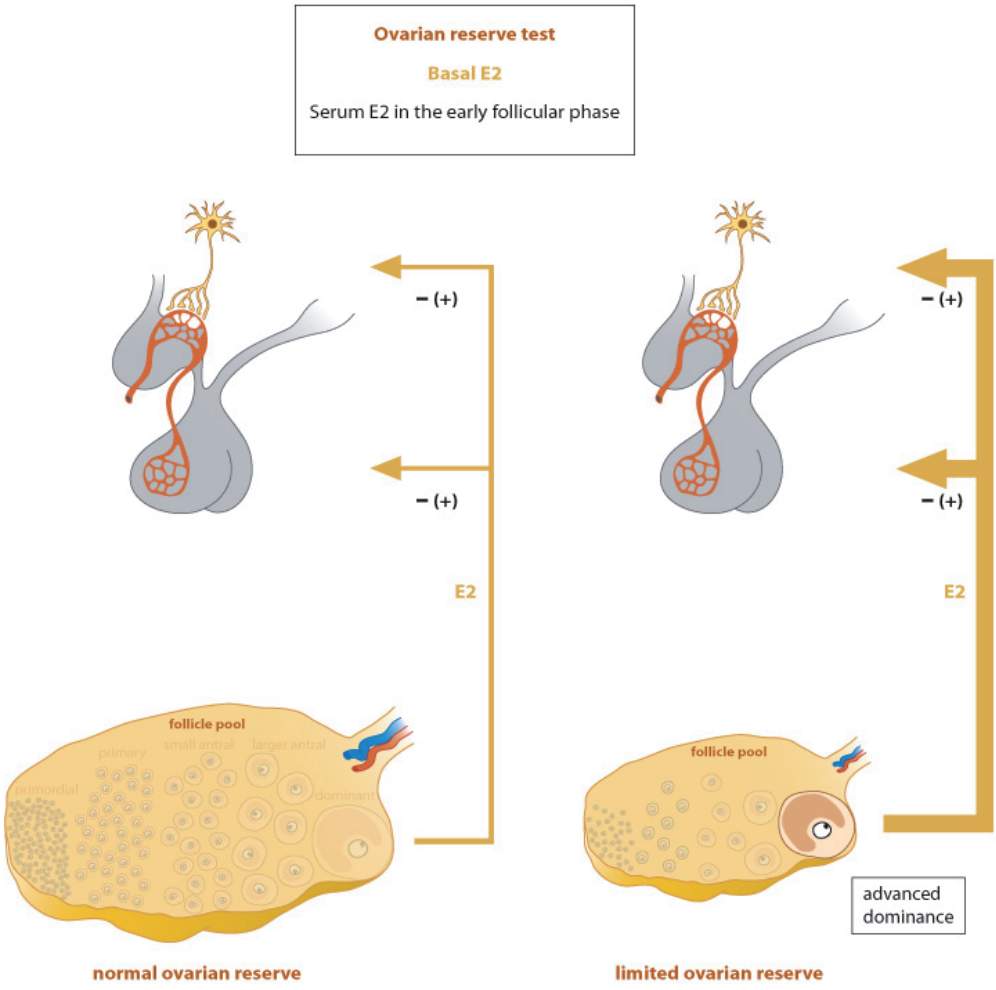


Basal estradiol

The pituitary FSH release in women of advancing reproductive age is less restrained than in young women by factors from the granulosa cell-oocyte complex. In these women, loss of inhibin-mediated negative feedback on the pituitary will result in increased FSH

output, starting already in the luteal phase, which may result in early follicle growth, with a consequent early production of estradiol (E2). This premature E2 elevation signifies early recruitment and is a common perimenopausal pattern (fig. 3). Elevated E2 levels exert a negative feedback on the hypothalamic-pituitary axis, reducing FSH secretion. Thus, on cycle day 3, a seemingly normal FSH level is accompanied by a high E2 level. The elevated basal estradiol level therefore is a negative prognostic indicator of response to stimulation in IVF patients with normal bFSH levels (Evers *et al.*, 1998, Smotrich *et al.*, 1995, Licciardi *et al.*, 1995). Some studies therefore advocate the use of the combined assessment of basal FSH and estradiol for the prediction of diminished ovarian reserve. However, the limited number of studies addressing this combined test prohibits a formal systematic review.

Figure 3. The principles of the ovarian reserve test: bestradiol



Basal inhibin B

The inhibins are dimeric polypeptides, including inhibin A and inhibin B. Both are believed to be granulosa cell products, with inhibin A (Groome *et al.*, 1994, Groome *et al.*, 1996, Klein *et al.*, 1996, Muttukrishna *et al.*, 1994) being secreted predominantly in the late follicular and luteal phase by the dominant follicle and corpus luteum, and inhibin B secreted predominantly in the follicular phase (Hall *et al.*, 1999) by the developing cohort of antral follicles in the cycle. Both inhibin A and inhibin B have the capacity to suppress FSH secretion by the pituitary without affecting LH secretion. Inhibins may also have paracrine functions influencing folliculogenesis in the ovary itself (Findlay, 1986a, Findlay, 1994, Findlay, 1993, Hillier, 1991).

Inhibin B concentrations are believed to provide a more direct assessment of ovarian reserve as inhibin B is mainly produced by the small antral follicles that constitute the FSH sensitive cohort. A decrease in inhibin B secretion and to some extent also inhibin A, as a result of a reduction in cohort size with ageing is generally held responsible for elevated FSH levels (fig. 4) and is associated with decreased oocyte quality and fertility potential (Bancsi *et al.*, 2002, Burger 1993, Burger *et al.*, 1995, de Koning *et al.*, 2000, Klein *et al.*, 1996, Seifer *et al.*, 1997, Seifer *et al.*, 1999). Especially in IVF, initial studies reported an association between diminished ovarian response and lower pregnancy rates on the one hand and decreased inhibin B levels on the other. Later studies have shed some doubt upon these findings (Hall *et al.*, 1999).

Basal volume of the ovaries

Real time two-dimensional (2D) pelvic ultrasonography is a relatively accurate and reliable method of determining ovarian volume and morphology (Campbell *et al.*, 1982). Interobserver and intraobserver measurements have been shown to be very low when using transvaginal sonography (Higgins *et al.*, 1990).

The mean ovarian volume increases from 0,7 ml at 10 years of age to 5,8 ml at 17 years of age (Ivarsson *et al.*, 1983). It has been suggested that there are no major changes in ovarian volume during reproductive years until the premenopausal period. In women > 40 years old, there is a dramatic drop in ovarian volume which is not related to parity (Higgins *et al.*, 1990, Ivarsson *et al.*, 1983, Andolf *et al.*, 1987). Thereafter there is a further sharp decline in size in postmenopausal women which seems mostly related to the time when menstruation ceases, rather than merely to age, because when oestrogen treatments were given, there appeared to be no decrease in ovarian volume with age (Andolf *et al.*, 1987).

Several studies (Syrop *et al.*, 1995, Tomás *et al.*, 1997, Lass *et al.*, 1997b, Bancsi *et al.*, 2002) demonstrate that ovarian volume, as determined by transvaginal ultrasonography, is a predictor of ovarian reserve and clinical pregnancy rate (fig. 5). The volume of the ovaries is an indirect indicator of the activity of the ovaries. The ovarian volume has a prognostic value in ovarian reserve and ovarian volume is correlated with the number of early antral follicles measured sonographically on cycle day 3.

Figure 4. The principles of the ovarian reserve test: binhibin B

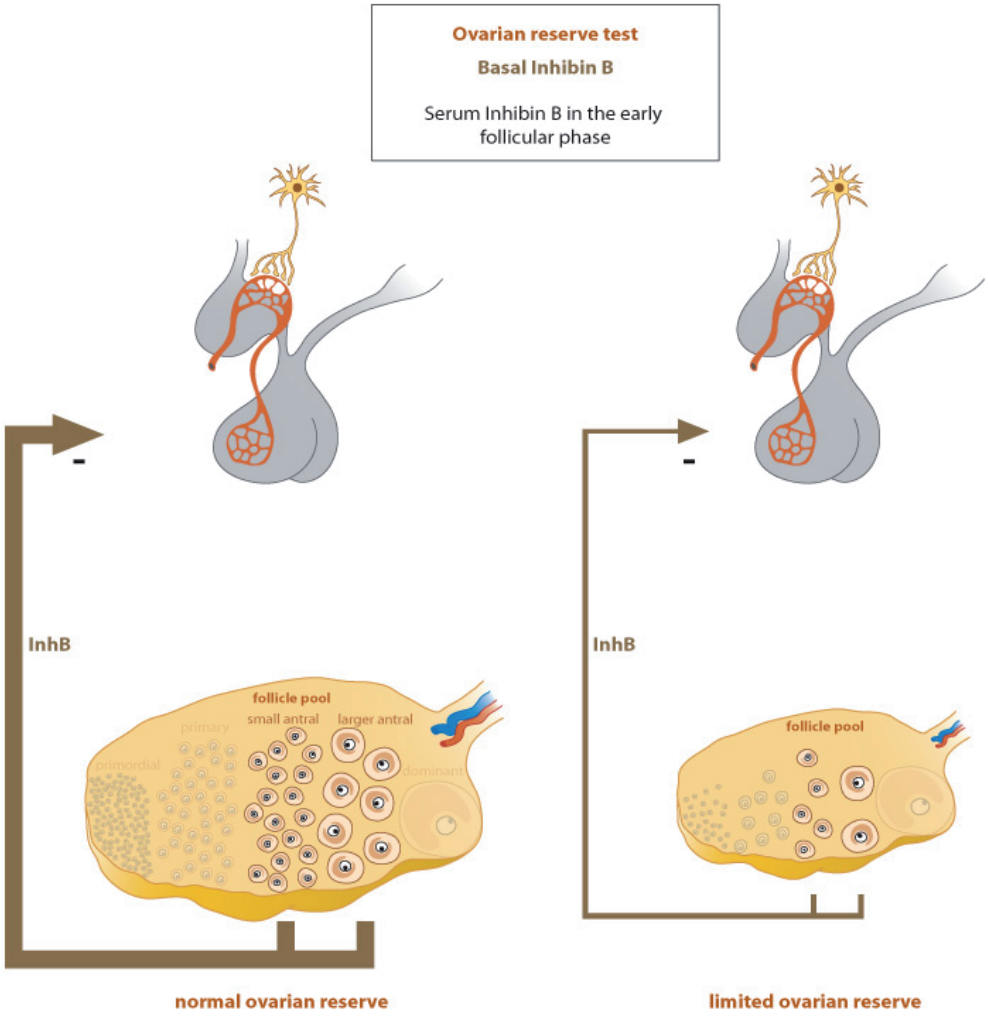
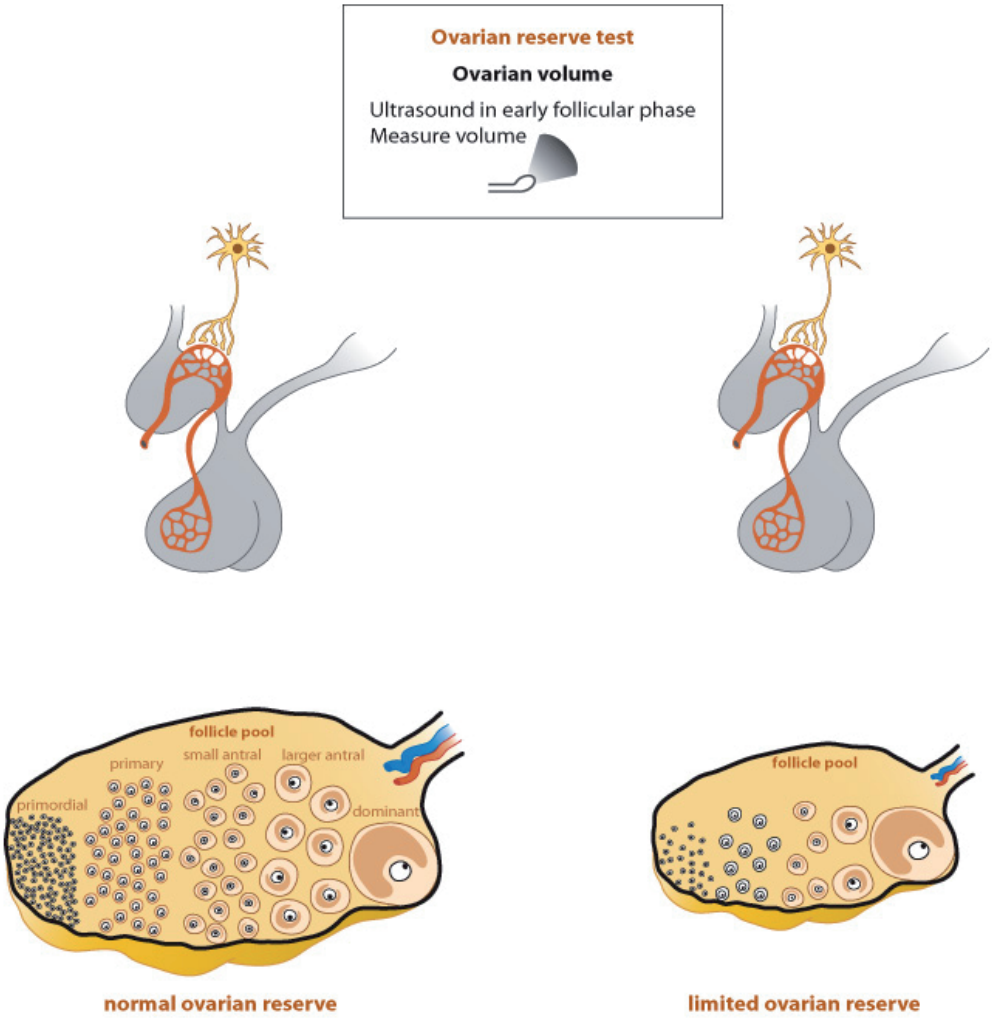


Figure 5. The principles of the ovarian reserve test: ovarian volume

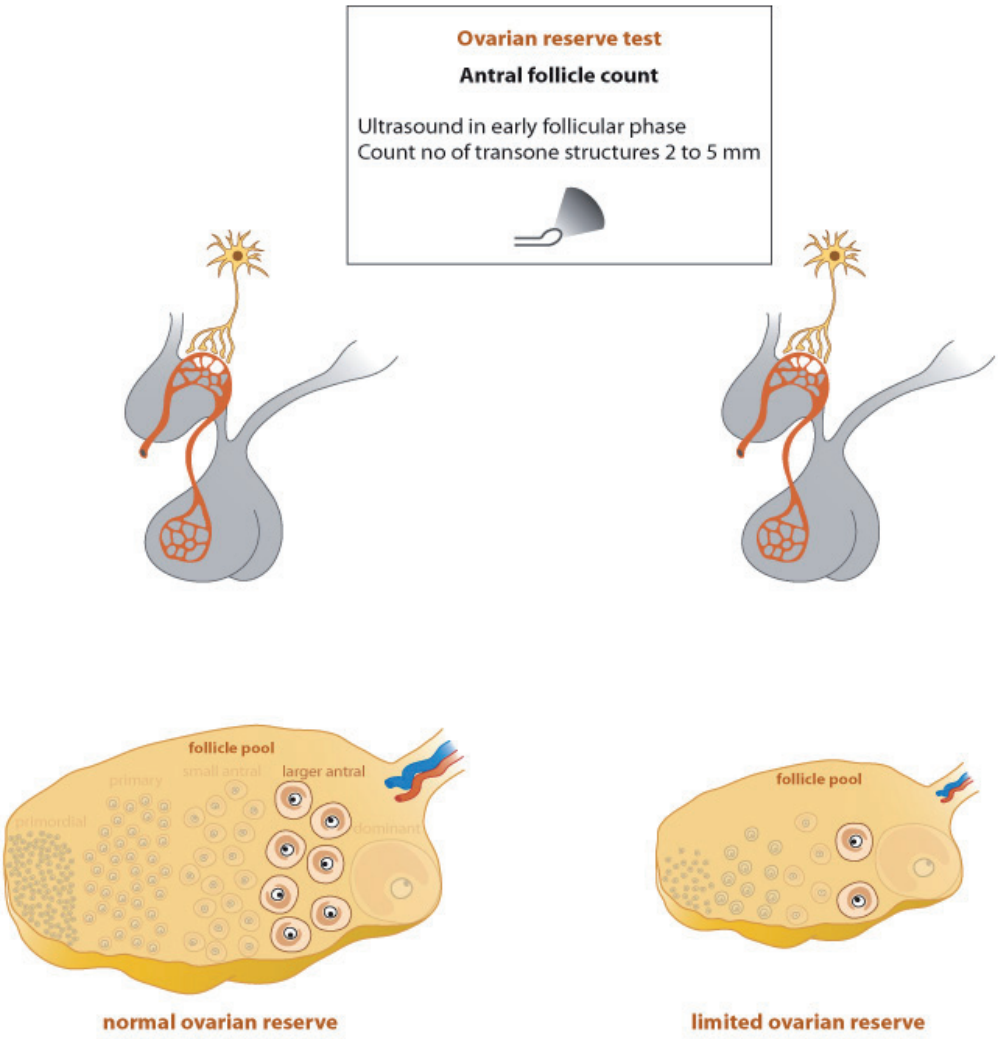


Basal count of the antral follicles

Ageing of the ovary is related to a gradual reduction in the number of primordial follicles (te Velde and Pearson, 2002). The number of primordial follicles declines exponentially through childhood and adult life leading to ovaries that are almost devoid of follicles at the age of menopause (Faddy, 2000, Faddy *et al.*, 1992a, Faddy *et al.*, 1992b). The number of primordial follicles in the ovary appears to be correlated with the number of antral follicles in all stages as shown by histological analysis (Gougeon, 1984). Also, with female ageing the decline in primordial follicle numbers parallels the decrease in size of the so called FSH sensitive, antral follicle cohort (Scheffer *et al.*, 1999). This is the number of follicles that has formed an antrum and varies in size between 2 and 10 mm (fig. 6). At any moment in the menstrual cycle this cohort is present as a result of a continuous process of supply from earlier follicle

stages and wastage by apoptosis. Assessment of the size of this antral follicle cohort present in both ovaries is possible by the use of transvaginal ultrasound (Meldrum *et al.*, 1984, Pache *et al.*, 1990) with favourable intra- and interobserver variability (Bancsi *et al.*, 2004a, Pache *et al*., 1990, Hansen *et al.*, 2003, Scheffer *et al.*, 2002). In recent years several papers have been published concerning the relation between the antral follicle count (AFC, defined as the total number of antral follicles, sized 2-5 or 2-10 mm, present in both ovaries) and the ovarian response in IVF (Bancsi *et al.*, 2002, Chang *et al.*, 1998a, Ng *et al.*, 2000), as well as the occurrence of the menopausal transition (van Rooij *et al.*, 2004), indicating that this parameter relates strongly to the quantitative aspects of ovarian reserve.

Figure 6. The principles of the ovarian reserve test: antral follicle count



Ovarian vascular flow

Adequate ovarian blood flow is an important precondition for normal physiological ovarian function (Findlay, 1986b, Reynolds *et al.*, 2002, Redmer, 1996) (fig. 7). Transvaginal Doppler ultrasound has made feasible the non-invasive evaluation of ovarian stromal blood flow during the menstrual and IVF stimulated cycles (Bassil *et al.*, 1997, Zaidi *et al.*, 1996a, Zaidi *et al.*, 1996b). Oocyte quality is sensitive to hypoxic damage (Van Blerkom, 2000, Van Blerkom *et al.*, 1997). Reduced oxygen delivery may result from increased resistances at the level of the perifollicular arteries (Battaglia *et al.*, 2000). The compromised perifollicular microcirculation and consequent hypoxia may cause an increased incidence of aneuploidal oocytes (Van Blerkom *et al.*, 1997). Estrogen has been shown to exert an important direct effect on the vascular wall, allowing for vasodilatation and improving blood flow (Bergqvist *et al.*, 1993). Estrogen receptors are found in human ovaries (Pelletier *et al.*, 2000, Taylor *et al.*, 2000) and in the vessel wall (Bergqvist *et al.*, 1993). Furthermore, estrogen can initiate the release of nitric oxide, a potent vasodilator, where it produces vasodilatation and improves blood flow. As a consequence of an increased ovarian steroidogenic tissue volume, a higher output of estrogens, progestins and vasoactive substances (vascular endothelial growth factor) is achieved (Engmann *et al.*, 1999). The rapid and exaggerated cyst formation in the ovaries requires angiogenesis, which leads to neovascularization. The indirect assessment of the degree of angiogenesis may be achieved by a power Doppler examination of the blood flow characteristics of the ovary (fig. 7). Reduced oxygen delivery may result from increased resistance at the level of the perifollicular arteries, as observed in poor responders.

Clomiphene Citrate Challenge Test (CCCT)

Clomiphene citrate challenge tests (CCCT) have been shown to be of predictive value for the ovarian response with respect to hyperstimulation and pregnancy rates in patients undergoing assisted reproduction. The CCCT also has a predictive value for the prognosis of pregnancy in a general infertility population (Scott *et al.*, 1993a).

Navot *et al.* (1987) described the Clomiphene citrate challenge test (CCCT) as a parameter for the ovarian reserve. The CCCT measures the bFSH on cycle day 3 and CD 10 after administration of 100 mg clomiphene citrate per day from CD 5 to CD 9. This test was modified by several other investigators (Loumaye *et al.*, 1990, Scott *et al.*, 1993b). The CCCT is an indirect test, because the ovaries are not stimulated directly, but through stimulation of the hypothalamic pituitary axis. This test assesses to what extent a growing cohort of follicles has the potency to generate and to maintain the negative feedback, by the release of inhibin B and, to a lesser extent, estradiol on the hypothalamic pituitary axis, leading to serum FSH-levels between certain limits (fig. 8). In all studies a significant difference was found in the pregnancy rates between groups of patients with a normal and abnormal CCCT using assisted reproduction techniques (Scott *et al.*, 1993a, Scott *et al.*, 1993b, Navot *et al.*, 1987, Loumaye *et al.*, 1990). The diminished ovarian reserve with increasing age starts off with a latent presence of ovarian failure: a normal bFSH, but an abnormal response to the CCCT. In a later phase occult ovarian failure develops, characterized by slightly elevated bFSH values, before menopause appears.

Hannoun *et al.* (1998) document the variation of the results of the CCCT performed in the same patient from cycle to cyle. They show a high degree of intercycle variability. Knowing the intercycle variability of an ovarian rererve test is very important for a correct interpretation of the results. Further research is required into whether repeated tests will improve the predictive value of the ovarian reserve test.

Figure 7. The principles of the ovarian reserve test: ovarian blood flow

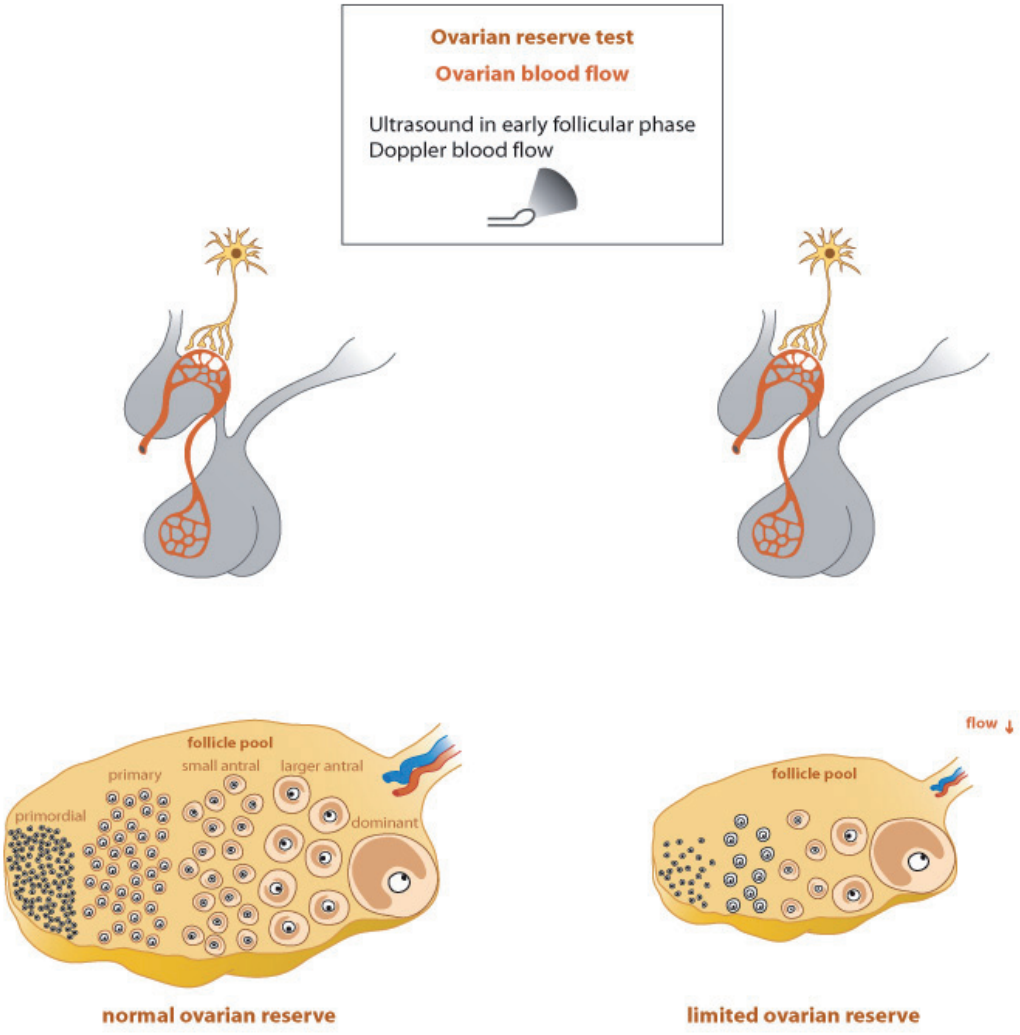
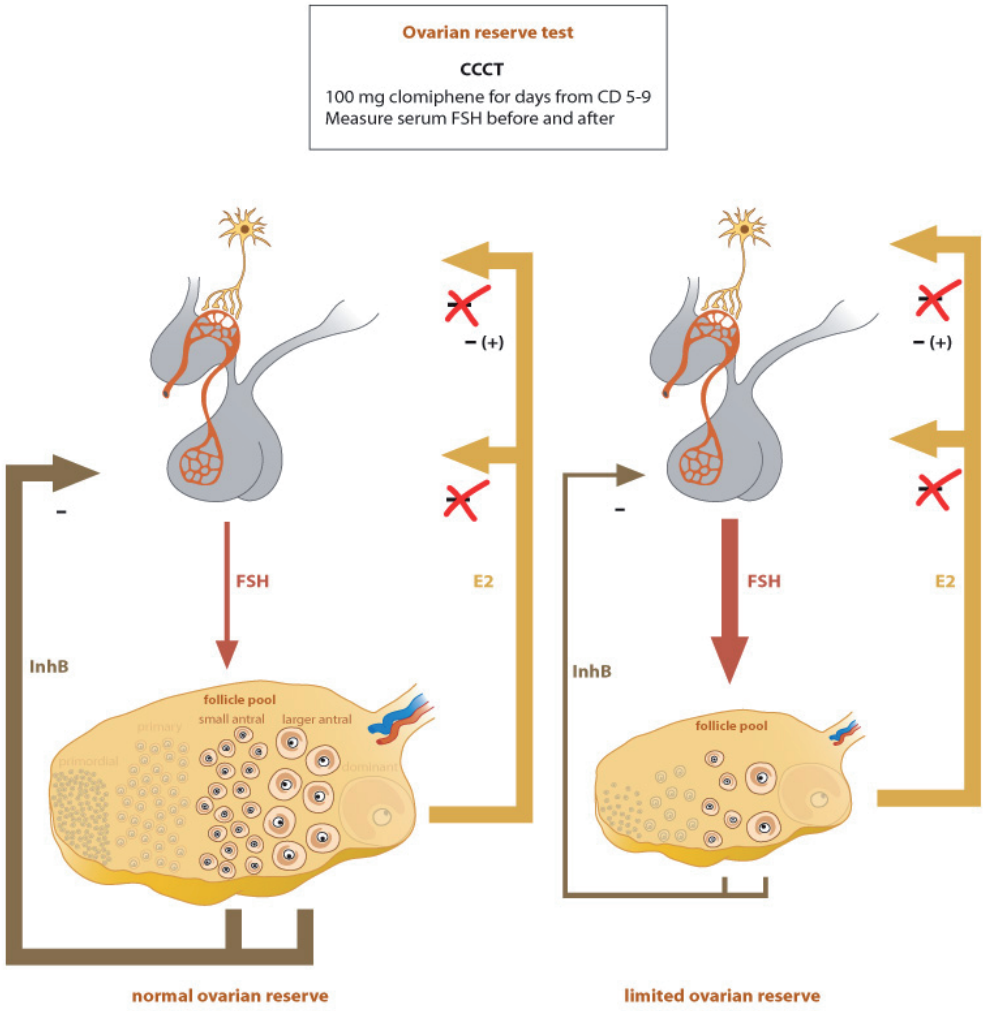


Figure 8. The principles of the ovarian reserve test: Clomiphene Citrate Challenge Test

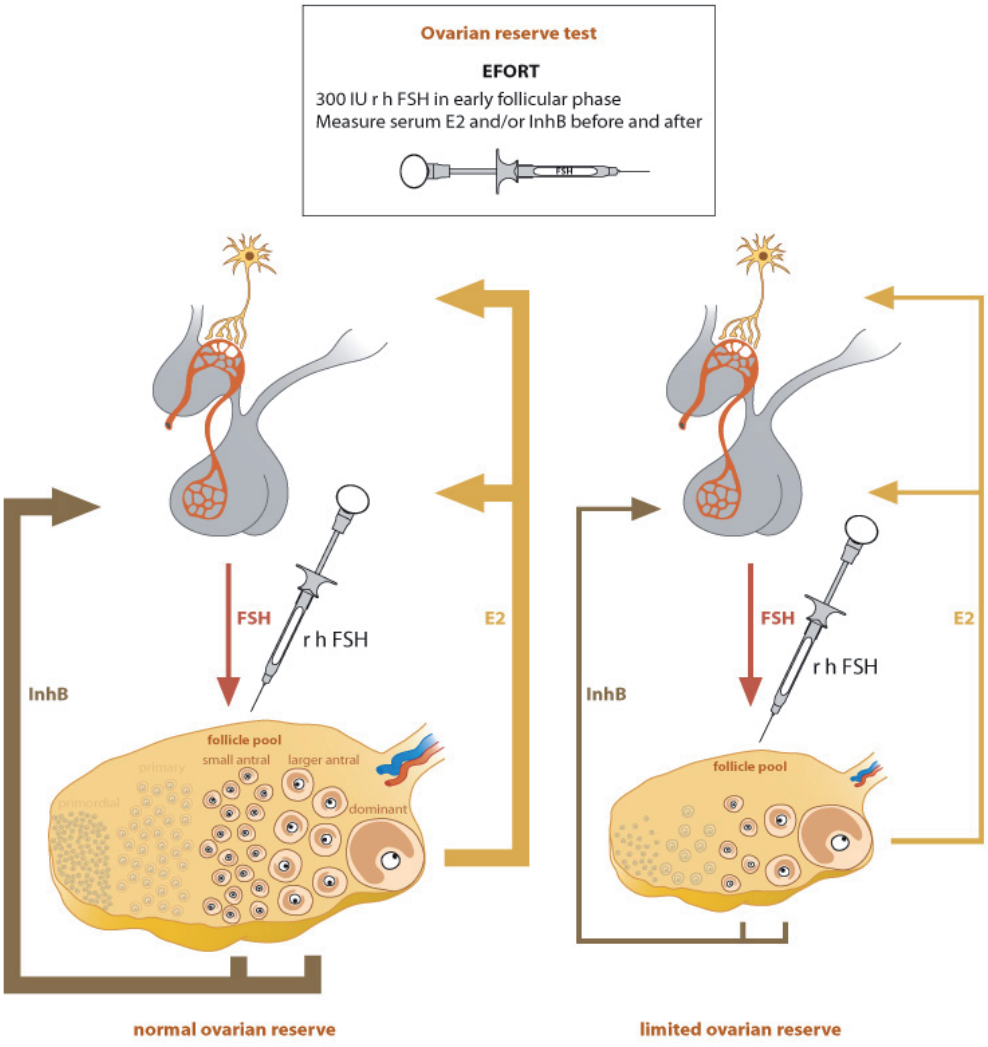


Exogenous FSH Ovarian Reserve Test (EFORT)

Fanchin *et al.* (1994) described the Exogenous FSH Ovarian Reserve Test (EFORT), as a test which can predict the ovarian reserve i.e. the subsequent ovarian response to stimulation for IVF and the pregnancy outcome. The EFORT adds a dynamic factor to the bFSH. This test demonstrates the ability of the ovaries to respond to a fixed dose of exogenously administered FSH (300 IU FSH) on cyle day 3 in 24 hours (fig. 9). FSH induces aromatase activity. The aromatase activity results in increased follicular concentrations of estradiol since the aromatase substrate, androstenedione is abundantly available. Thereby the estradiol increase becomes a parameter for the size of a growing cohort. The two parameters evaluated in this test are: The increase of serum E2 from cycle day 3 till cycle day 4 (= δE2), in combination with the bFSH. The outcome of ovarian hyperstimulation is classified as a normal, moderate or poor response depending on the following parameters: 1. the number of ampoules of FSH required to achieve ovarian stimulation, 2. the serum E2 level determined on the day of

hCG administration and 3. the number of retrieved mature oocytes. Fanchin *et al.* found a sensitivity of 90 % and a specificity of 81 % for the EFORT, while the values for the bFSH alone being 60 % and 45 % respectively.
In 2000, Dzik *et al.* (2000) suggested that an increase of serum inhibin B in the EFORT might be a better predictor for ovarian reserve than the E2-increment in the EFORT. Further research is necessary.

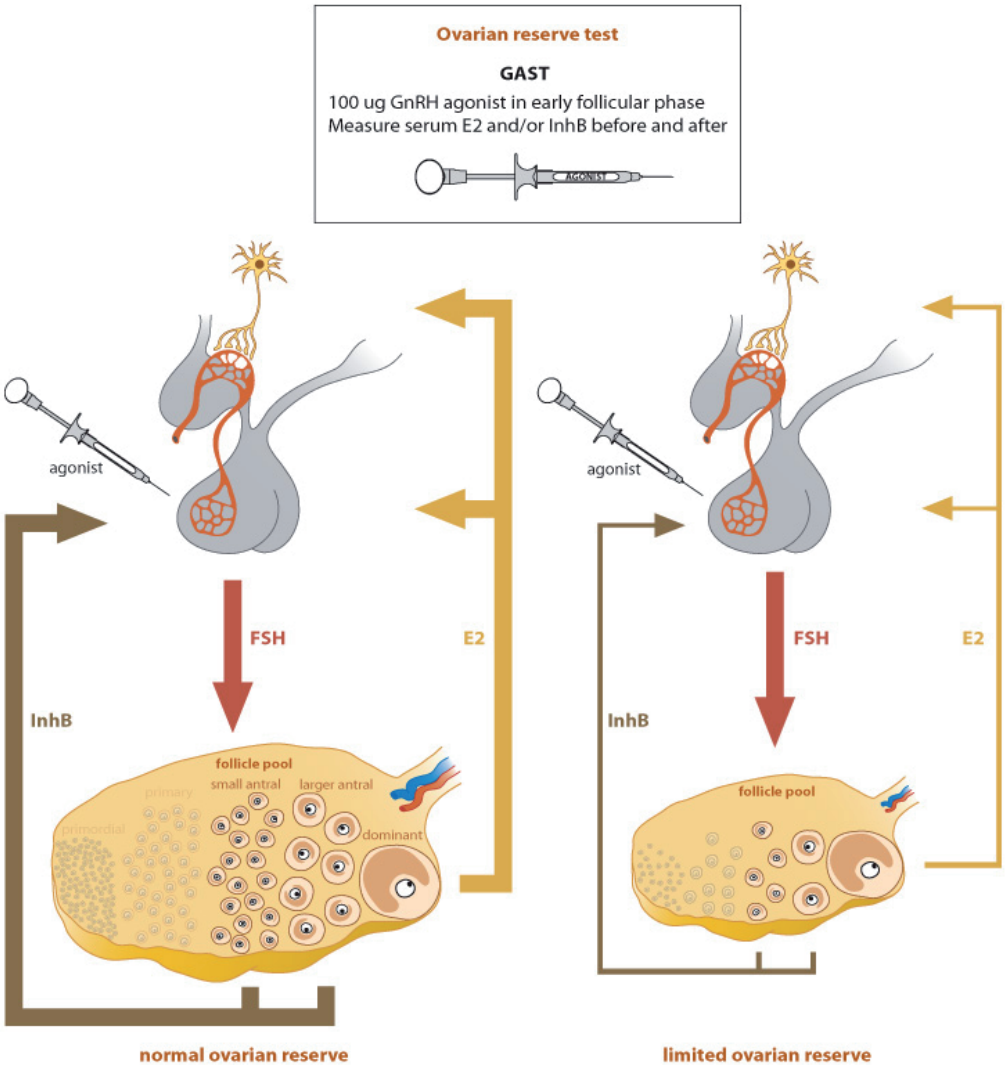
Figure 9. The principles of the ovarian reserve test: Exogenous FSH Ovarian Reserve Test



Gonadotropin Agonist Stimulation Test (GAST)
By subcutaneously administering a standard, supraphysiological, dose of a GnRH agonist a massive pituitary response is elicited; leading to the sustained elevation of LH and FSH levels for a period of at least 24 hours. The continuous exposure of the ovary to high levels

of FSH urges FSH sensitive antral follicles to increase the production of estradiol and Inhibin B in such a way that the rise from baseline is an expression of the size of the antral follicle cohort in the early follicular phase of the cycle (fig 10). Although the bioavailability of the GnRH agonist is approximately 100%, the pituitary response across individuals may vary. Still, even in relatively low responders the threshold for FSH dependent follicular growth is surpassed greatly and maximal ovarian stimulation will therefore be present in every individual (Broekmans *et al.*, 1993). The GAST is thereby a test that typically measures the quantitative aspects of ovarian reserve and describes the gradual decline in ovarian reserve with ageing (Scheffer *et al.*, 2003)

Figure 10. The principles of the ovarian reserve test: Gonadotrophin Agonist Stimulation Test



Multivariate models

Combining several known ovarian reserve tests introduces the possibility to refine the approach of estimating ovarian reserve. Most tests that are used in combination focus on the quantity aspect of ovarian reserve, like the AFC, AMH, FSH and Inhibin B. Tests that more easily represent oocyte quality like female age may well be incorporated into such a model, especially as the result of this test is always available. The presumption that every test may describe the ovarian reserve status from a slightly different perspective allows these tests to provide predictive information on top of that of others and as such to improve the overall predictive capacity. In addition, in any study on ovarian reserve testing where multifactor models were used, it can be assumed that these models represents the best predictive system for that population. By reviewing these test models in a meta-analytic fashion it can be assessed whether multifactor models in general will provide better test properties than single tests. If so, predictive models may be developed within institutions that perform best for daily practice, although not uniformly applicable.

Epilogue

Assisted reproduction techniques (ART) are complex and expensive and have strict indications. Informing infertile couples about their chances of pregnancy, spontaneously or by means of ART should have a high priority. Several factors play a significant role in determining the success rate of ART. One of the most important factors is the number of embryos available for embryo transfer (Templeton *et al.*, 1996). Fertilization is determined on one hand by the number and quality of the oocytes obtained by follicular aspiration after maximal ovarian hyperstimulation and on the other hand by the fertilizing capacity of the semen. An adequate dose of gonadotropins doesn't always result in a suitable ovarian response. In case of a poor response, very few or even no follicles develop, resulting in cancellation or if the treatment cycle is completed the chance of pregnancy is reduced. An exaggerated ovarian response can be dangerous leading to the development of many follicles and subsequently very high serum E2 levels. This situation can also lead to cancellation because of the risk of an ovarian hyperstimulation syndrome (OHSS). This clinical condition may be life threatening. The identification of certain patients before the start of an expensive, time consuming and often stressful IVF-treatment cycle is of major importance. It would enable clinicians to adjust the ovarian stimulation protocol on an individual basis. This policy may lead to an optimal chance for every couple, and as such may reduce the number of cancelled ART-treatment cycles because of poor or exaggerated ovarian responses. This information is best used for counseling patients regarding their individual chances for pregnancy. If women in an IVF-population have diminished ovarian reserve, they should be informed about the options of oocyte donation or adoption. What is needed is a simple, minimally invasive, reliable ovarian reserve test, which can predict the cohort of small antral follicles in the early follicular phase from which follicles can be recruited. After a survey through the literature, which test or combination of tests could predict the ovarian reserve, we came to the conclusion that there are several static and dynamic tests for ovarian reserve available, of which some predict reasonably well a poor or adequate response upon ovarian hyperstimulation and a prognosis for pregnancy, but did not give the prediction of the precise size of the cohort. None of these tests were in fact developed for determination of ovarian reserve. Also they all have a number of pitfalls in interpreting the results. It is important to know in what kind of population the test is used, because the outcome of the test is dependent on the prevalence of infertility for that

group. When the same test, with the same sensitivity and specificity, is used in a population with a low prevalence, the positive predictive value would be much lower (Barnhart *et al.*, 1999). Not all tests are used in the same populations and especially not in the general infertility population. There are few data documenting the variation of the ovarian reserve tests performed in the same patient from cycle to cycle. Knowing the intercycle variability of an ovarian stress test is very important for a good interpretation of the test.
For the interpretation of the test one must be aware of the differences between the assays and the laboratory. Everyone should define their own cut off values.
The ultimate goal is to offer a patient a tailor made individual treatment schedule and inform patients about their chances of pregnancy. Further validation including reproducibility and intercycle variation is necessary.

Aim of the thesis

The aim of the studies described in this thesis was to find an answer to the following questions:

a. Which ovarian reserve test or a certain combination can predict the cohort size of small antral follicles in the early follicular phase.

b. Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of poor and hyper ovarian response in an IVF population.

c. Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of pregnancy in an IVF population.

We approached these questions in two ways.

1. A prospective study was conducted that compared in an integral way all currently available static ovarian reserve tests: early follicular phase blood values of follicle stimulating hormone (FSH), oestradiol (E2), inhibin B and anti-mullerian hormone (AMH), the dynamic ovarian reserve tests: the exogenous FSH ovarian reserve test (EFORT), the Clomiphene Citrate Challenge Test (CCCT), the ultrasound tests: antral follicle count (AFC), basal ovarian volume (BOV) and the intercycle variability of test results with regard to the prediction of the ovarian response after ovarian hyperstimulation in an IVF treatment. The results of this study are reported in chapters 2, 3, 4, 5 and 6.

2. A systematic review of the literature was provided including an a priori protocolised information retrieval on all currently available and applied tests, namely early follicular phase blood values of follicle stimulating hormone (FSH), oestradiol, inhibin B and anti-mullerian hormone (AMH), the antral follicle count (AFC), the ovarian volume and the ovarian blood flow and furthermore the clomiphene citrate challenge test (CCCT), the exogenous FSH ovarian reserve test (EFORT) and the gonadotropin releasing hormone agonist stimulation test (GAST) as measures to determine ovarian reserve and their capability to predict ovarian response and chance of pregnancy. This systematic review is reported in chapter 7.

References

Andolf E, Jörgensen C, Svalenius E, Sundén B. Ultrasound measurement of the ovarian volume. Acta Obstet Gynecol Scand 1987;66:387-89.

Bancsi LF, Broekmans FJ, Eijkemans MJ, de Jong FH, Habbema JD, te Velde ER. Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 2002;77:328-36.

Bancsi LF, Broekmans FJ, Mol BW, Habbema JD, te Velde ER. Performance of basal follicle-stimulating hormone in the prediction of poor ovarian response and failure to become pregnant after in vitro fertilization: a meta-analysis. Fertil Steril 2003;79:1091-1100.

Bancsi LF, Broekmans FJ, Looman CW et al. Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. Fertil Steril 2004a;81:35-41.

Barnhart K, Osheroff J. We are overinterpreting the predictive value of serum follicle-stimulating hormone levels. Fertil Steril 1999;72:8-9.

Bassil S, Wyns C, Toussaint Demylle D et al. The relationship between ovarian vascularity and the duration of stimulation in in-vitro fertilization. Hum Reprod 1997;12:1240-45.

Battaglia C, Genazzani AD, Regnani G et al. Perifollicular Doppler flow and follicular fluid vascular endothelial growth factor concentrations in poor responders. Fertil Steril 2000;74;809-12.

Beemsterboer SN, Homburg R, Gorter NA, Schats R, Hompes PG, Lambalk CB. The paradox of declining fertility but increasing twinning rates with advancing maternal age. Hum Reprod 2006;21:1531-2

Behringer RR, Finegold MJ, Cate RL. Mullerian-inhibiting substance function during mammalian sexual development. Cell 1994;79:415-5 .

Bergqvist A, Bergqvist D, Ferno M. Estrogen and progesterone receptors in vessel walls. Biochemical and immunochemical assays. Acta Obstet Gynecol Scand 1993 ;72:10-6.

Block E. Quantitative morphological investigations of the follicular system in women. Variations at different ages. Acta Anat 1952;14:108-23.

Broekmans FJ, Bernardus RE, Broeders A et al. Pituitary responsiveness after administration of a GnRH agonist depot formulation: Decapeptyl CR. Clin Endocrinol Oxf 1993;38:579-87.

Burger HG. Clinical review 46: Clinical utility of inhibin measurements. J Clin Endocrinol Metab 1993;76:1391-6.

Burger HG, Dudley EC, Hopper JL et al. The endocrinology of the menopausal transition: a cross-sectional study of a population-based sample. J Clin Endocrinol Metab 1995;80:3537-45.
Cahill DJ, Prosser CJ, Wardle PG, Ford WCL, Hull MGR. Relative influence of serum follicle stimulating

hormone, age and other factors on ovarian response to gonadotrophin stimulation. Br J Obstet Gynaecol 1994; 101:999-1002.

Campbell S, Goessens L, Goswamy R, Whitehead M. Real-time ultrasonography for determination of ovarian morphology and volume. A possible early screening test for ovarian cancer? Lancet 1982;i:425-6.

Chang MY, Chiang CH, Hsieh TT, Soong YK, Hsu KH. Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil Steril 1998;69:505-10.

Danforth DR, Arbogast LK, Mroueh J, Kim MH, Kennard EA, Seifer DB, et al. Dimeric inhibin: a direct marker of ovarian aging. Fertil Steril 1998;70:119-23.

De Koning CH, Popp-Snijders C, Schoemaker J et al. Elevated FSH concentrations in imminent ovarian failure are associated with higher FSH and LH pulse amplitude and response to GnRH. Hum Reprod 2000;15:1452-6.

Durlinger AL, Kramer P, Karels B, de Jong FH, Uilenbroek JT, Grootegoed JA, Themmen AP. Control of primordial follicle recruitment by anti-Mullerian hormone in the mouse ovary. Endocrinology 1999;140:5789-96.

Durlinger AL, Gruijters MJ, Kramer P, Karels B, Kumar TR, Matzuk MM, Rose UM, de Jing FH, Uilenbroek JT, Grootegoed JA, Themmen AP. Anti-Mullerian hormone attenuates the effects of FSH on follicle development in the mouse ovary. Endocrinology 2001;142:4891-9.

Dzik A, Lambert-Messerlian G, Izzo VM, Soares JB, Pinotti JA, Seifer DB. Inhibin B response to EFORT is associated with the outcome of oocyte retrieval in the subsequent in vitro fertilization cycle. Fertil Steril 2000;74:1114-7.

Engmann L, Sladkevicius P, Agrawal R et al. Value of ovarian stromal blood flow velocity measurement after pituitary suppression in the prediction of ovarian responsiveness and outcome of in vitro fertilization treatment. Fertil Steril 1999;71:22-9.

Evers JL, Slaats P, Land JA, Dumoulin JCM, Dunselman GAJ. Elevated levels of basal estradiol-17□ predict poor response in patients with normal basal levels of follicle-stimulating hormone undergoing in vitro fertilization. Fertil Steril 1998;69:1010-4.

Faddy MJ. Follicle dynamics during ovarian ageing. Mol Cell Endocrinol 2000;163:43-8.

Faddy MJ, Gosden RG, Gougeon A et al. Accelerated disappearance of ovarian follicles in mid-life: implications for forecasting menopause. Hum Reprod 1992a;7:1342-46.

Faddy MJ, Gosden RG, Gougeon A et al. Accelerated disappearance of ovarian follicles in mid-life: implications for forecasting menopause. Hum Reprod 1992b;7:1342-6.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A, Frydman R. Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Human Reprod 1994;9:1607-11.

Fanchin R, Schonauer LM, Righini C, frydman N, Frydman R, Taieb J. Serum anti-Mullerian hormone dynamics during controlled ovarian hyperstimulation. Hum Reprod 2003;18:328-32.

Findlay JK. The nature of inhibin and its use in the regulation of fertility and diagnosis of infertility. Fertil Steril 1986a;46:770-783.

Findlay JK. Angiogenesis in reproductive tissues. J Endocrinol 1986b;111:357-66.

Findlay JK. An update on the roles of inhibin, activin, and follistatin as local regulators of folliculogenesis. Biol Reprod 1993;48:15-23.

Findlay JK. Peripheral and local regulators of folliculogenesis. Reprod Fertil Dev 1994;6:127-139.

Gougeon A. Caracteres qualitatifs et quantitatifs de la population folliculaire dans l'ovaire humaine adulte. Contracept. Fertil.Sex 1994 ;12:527-35.

Groome NP, Illingworth PJ, O'Brien M et al. Detection of dimeric inhibin throughout the human menstrual cycle by two-site enzyme immunoassay. Clin Endocrinol Oxf 1994;40:717-23.

Groome NP, Illingworth PJ, O'Brien M, Pai R, Rodger FE, Mather JP, et al. Measurement of dimeric inhibin B throughout the human menstrual cycle. J Clin Endocrinol Metab 1996;81:1401-5.

Gülekli B, Bulbul Y, Onvural A, Yorukoglu K, Posaci C, Demir N. Accuracy of ovarian reserve tests. Hum Reprod 1999;14:2822-6.

Hall JE, Welt CK, Cramer DW. Inhibin A and inhibin B reflect ovarian function in assisted reproduction but are less useful at predicting outcome. Hum Reprod 1999;14:409-15.

Hannoun A, Abu Musa A, Awwad J, Kaspar H, Khalil A. Clomiphene citrate challenge test: cycle to cycle variability of cycle day 10 follicle stimulating hormone level. Clin Exp Obstet Gyn 1998;25:155-6.

Hansen LM, Batzer FR, Gutmann JN, Corson SL, Kelly MP, Gocial B. Evaluating ovarian reserve: follicle stimulating hormone and oestradiol variability during cycle days 2-5. Hum Reprod 1996;11:486-9.

Hansen KR, Morris JL, Thyer AC et al. Reproductive aging and variability in the ovarian antral follicle count: application in the clinical setting. Fertil Steril 2003;80:577-83.

Hehenkamp JK, Loomans CWN, Themmen APN, de Jong FH, te Velde ER, Broekmans FJM. Anti-Mullerian Hormone levels in the spontaneous menstrual ccle do not show substantial fluctuation. J Clin Endocrinol Metab 2006;10:4057-63.

Higgins RV, van Nagell JR, Woods CH, Thompson EA, Kryscio RJ. Interobserver variation in ovarian measurements using transvaginal sonography. Gynecol Oncol 1990;39:69-71.

Hillier SG. Paracrine Control of Follicular Estrogen Synthesis. Sem Reprod Endocrinol 1991;9:332-40.

Hughes EG, King C, Wood EC. A prospective study of prognostic factors in in vitro fertilization and embryo transfer. Fertil Steril 1989;51:838-44.

Ivarsson SA, Nillson KO, Persson PH. Ultrasonography of the pelvic organs in prepubertal and postpubertal girls. Arch Dis Child 1983;58:352-4.

Khalifa E, Toner JP, Muasher SJ, Acosta AA. Significance of basal follicle-stimulating hormone levels in women with one ovary in a program of in vitro fertilization. Fertil Steril 1992;57:835-9.

Klein NA, Illingworth PJ, Groome NP, McNeilly AS, Battaglia DE, Soules MR. Decreased Inhibin B secretion is Associated with the monotropic FSH rise in older, ovulatory women: A study of serum and follicular fluid levels of dimeric inhibin A and B in spontaneous menstrual cycles. J Clin Endocrinol Metab 1996;81:2742-5.

Lambalk CB, Boomsma DI, de Boer L, de Koning CH, Schoute E, Popp-Snijders C, et al. Increased levels and pulsatility of follicle-stimulating hormone in mothers of hereditary dizygotic twins. J Clin Endocrinol Metab 1998;83:481-6.

Lambalk CB, de Koning CH, Flett A, van Kasteren Y, Gosden R, Homburg R. Assessment of ovarian reserve. Ovarian biopsy is not a valid method for the prediction of ovarian reserve. Hum Reprod 2004;19: 1055-1059.

Lass A, Silye R, Abrams D C, Krausz T, Hovatta O, Margara R, Winston R M. Follicular density in ovarian biopsy of infertile women: a novel method to assess ovarian reserve. Hum Reprod 1997a;12:1028-1031.

Lass A, Skull J, McVeigh E, Margara R, Winston RML. Measurement of ovarian volume by transvaginal sonography before ovulation induction with human menopausal gonadotrophin for in-vitro fertilization can predict poor response. Hum Reprod 1997b;12:294-7.

Lass A. Assessment of ovarian reserve - is there a role for ovarian biopsy? Hum Reprod 2001;16:1055-7.

Lass A. Assessment of ovarian reserve: is there still a role for ovarian biopsy in the light of new data? Hum Reprod 2004;19:467-469.

Lenton EA, Sexton L, Lee S, Cooke ID. Progressive changes in LH and FSH and LH: FSH ratio in women throughout reproductive life. Maturitas 1988;10:35-43.

Lenton EA, Landgren BM, Sexton L, Harper R. Normal variation in the length of the follicular phase of the menstrual cycle: effect of chronological age. Br J Obstet Gynaecol 1984;91:681-4.

Licciardi FL, Liu HC, Rosenwaks Z. Day 3 estradiol serum concentrations as prognosticators of ovarian stimulation respons and pregnancy outcome in patients undergoing in vitro fertilization. Fertil Steril 1995;64:991-4.

Loumaye E, Billion JM, Mine JM, Psalti I, Pensis M, Thomas K. Prediction of individual response to controlled ovarian hyperstimulation by means of a clomiphene citrate challenge test. Fertil Steril 1990;53:295-301.

Martin JSB, Nisker JA, Tummon IS, Daniel SAJ, Auckland JL, Feyles V. Future in vitro fertilization pregnancy potential of women with variably elevated day 3 follicle-stimulating hormone levels. Fertil Steril 1996;65:1238-40.

Meldrum DR. Female reproductive aging-ovarian and uterine factors. Fertil Steril 1993;59:1-5.

Muttukrishna S, Fowler PA, Groome NP et al. Serum concentrations of dimeric inhibin during the spontaneous human menstrual cycle and after treatment with exogenous gonadotrophin. Hum Reprod 1994;9:1634-42.

Navot D, Rosenwaks Z, Margalioth EJ. Prognostic assessment of female fecundity. Lancet 1987;2:645-7.

Navot D, Drews MR, Bergh PA, Guzman I, Kaerstaedt A, Scott RT, et al. Age-related decline in female fertility is not due to dimished capacity of the uterus to sustain embryo implantation. Fertil Steril 1994;61;97-101.

Ng EH, Tang OS, Ho PC. The significance of the number of antral follicles prior to stimulation in predicting ovarian responses in an IVF programme. Hum Reprod 2000;15:1937-42.

Pache TD, Wladimiroff JW, de Jong FH et al. Growth patterns of nondominant ovarian follicles during the normal menstrual cycle. Fertil Steril 1990;54:638-42.

Pearlstone AC, Fournet N, Gambone JC, Pang SC, Buyalos RP. Ovulation induction in women age 40 and older: the importance of basal follicle-stimulating hormone level and chronological age. Fertil Steril 1992;58: 674-9.

Pelletier G and El Alfy M. Immunocytochemical localization of estrogen receptors alpha and beta in the human reproductive organs. J Clin Endocrinol Metab 2000;85:4835-40.

Qu J, Godin PA, Nisolle M et al. Distribution and epidermal growth factor receptor expression of primordial follicles in human ovarian tissue before and after cryopreservation. Hum Reprod 2000;15:302-10.

Redmer DA and Reynolds LP. Angiogenesis in the ovary. Rev Reprod 1996;1:182-92.

Reynolds LP, Grazul-Bilska AT, Redmer DA. Angiogenesis in the female reproductive organs: pathological implications. Int Jexp Pathol 2002;83:151-63.

Sharara F I, Scott R T. Assessment of ovarian reserve. Is there still a role for ovarian biopsy? First do no harm! Hum Reprod 2004;19:470-1.

Scheffer GJ, Broekmans FJ, Dorland M, Habbema JD, Looman CW , te Velde ER. Antral follicle counts by transvaginal ultrasonography are related to age in women with proven natural fertility. Fertil Steril 1999;72: 845–51.

Scheffer GJ, Broekmans FJ, Bancsi LF et al. Quantitative transvaginal two- and three-dimensional sonography of the ovaries: reproducibility of antral follicle counts. Ultrasound Obstet Gynecol 2002;20:270-5.

Scheffer GJ, Broekmans FJ, Looman CW et al. The number of antral follicles in normal women with proven fertility is the best reflection of reproductive age. Hum Reprod 2003;18:700-6.

Schmidt KL, Ernst E, Byskov AG et al. Survival of primordial follicles following prolonged transportation of ovarian tissue prior to cryopreservation. Hum Reprod 2003;18:2654-9.

Scott RT, Toner JP, Muasher SJ, Oehninger S, Robinson S, Rosenwaks Z. Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. Fertil Steril 1989;51:651-4.

Scott RT, Hofmann GE, Oehninger S, Muasher SJ. Intercycle variability of day 3 follicle-stimulating hormone levels and its effect on stimulation quality in in vitro fertilization. Fertil Steril 1990;54:297-302.

Scott RT, Leonardi MR, Hofmann GE, Illions EH, Neal GS, Navot D. A prospective evaluation of clo miphene citrate challenge test screening in the general infertility population. Obstet Gynecol 1993a;82:539-44.

Scott RT, Illions EH, Kost ER, Dellinger C, Hofmann GE, Navot D. Evaluation of the significance of the estradiol response during the clomiphene citrate challenge test. Fertil Steril 1993b;60:242-6.

Scott RT, Opsahl MS, Leonardi MR, Neall GS, Illions EH, Navot D. Life table analysis of pregnancy rates in a general infertility population relative to ovarian reserve and patient age. Hum Reprod 1995;10:1706-10.

Seifer DB, Lambert-Messerlian G, Hogan JW. Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome. Fertil Steril 1997;67:110-4.

Seifer DB, Scott RT, Bergh PA, Abrogast LK, Friedman CI, Mack CK et al. Women with declining ovarian reserve may demonstrate a decrease in day 3 serum inhibin B before a rise in day 3 follicle- stimulating hormone. Fertil Steril 1999;72:63-5.

Seifer DB, Mac Laughlin DT, Christian BP, Feng B, Shelden RM. Early follicular serum mullerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. Fertil Steril 2002;77:468-71.

Sherman BM, West JH, Korenman SG. The menopausal transition: Analysis of LH, FSH, estradiol, and progesterone concentrations during menstrual cycles of older women. J Clin Endocrinol Metab 1976;42:629-36.

Smotrich DB, Widra EA, Gindoff PR, Levy MJ, Hall JL, Stillman RJ. Prognostic value of day 3 estradiol on in vitro fertilization outcome. Fertil Steril 1995;64:1136-40.

Syrop CH, Willhoite A, Van Voorhis BJ. Ovarian volume: a novel outcome predictor for assisted reproduction. Fertl Steril 1995;64:1167-71.

Taylor AH and Al Azzawi F. Immunolocalisation of oestrogen receptor beta in human tissues. J Mol Endocrinol 2000;24:145-55.

Templeton A, Morris JK and Parslow W. Factors that affect outcome of in-vitro fertilisation treatment Lancet 1996;348:1402-6.

Te Velde ER and Pearson PL. The variability of female reproductive ageing. Hum Reprod Update 2002;8:141-54.

Tomás C, Nuojua-Huttunen S, Martikainen H. Pretreatment transvaginal ultrasound examination predicts ovarian responsiveness to gonadotrophins in in-vitro fertilization. Hum Reprod 1997;12:220-3.

Toner JP, Philput CB, Jones GS, Muasher SJ. Basal follicle-stimulating hormone level is a better predictor of in vitro fertilization performance than age. Fertil Steril 1991;55:784-91.

Toner JP. The significance of elevated FSH for reproductive function. Bail Clin Obstet Gynaecol 1993;7:283-95.

Treloar AE, Boynton RE, Behn BG, Brown BW. Variation of the human menstrual cycle through reproductive life. Int J Fertil 1967;12:77-126.

Van Blerkom J. Intrafollicular influences on human oocyte developmental competence: perifollicular vascularity, oocyte metabolism and mitochondrial function. Hum Reprod 2000;15:173-88.

Van Blerkom J, Antczak M, Schrader R. The developmental potential of the human oocyte is related to the dissolved oxygen content of follicular fluid: association with vascular endothelial growth factor levels and perifollicular blood flow characteristics. Hum Reprod 1997;12:1047-55.

Van Rooij IA, Broekmans FJ, te Velde ER, Fauser BC, Bancsi LF, de Jong FH, Themmen AP. Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod 2002;17:3065-71.

Van Rooij IA, Tonkelaar I, Broekmans FJ, Looman CW, Scheffer GJ, de Jong FH, Themmen AP, te Velde ER. Anti-mullerian hormone is a promising predictor for the occurrence of the menopausal transition. Menopause 2004;11:601-6.

Van Rooij IA, Broekmans FJ, Scheffer GJ, Looman CW, Habbema JD, de Jong FH, Fauser BJ, Themmen AP, te Velde ER. Serum antimullerian hormone levels best reflect the reproductive decline with age in normal women with proven fertility: A longitudinal study. Fertile Steril 2005; 83:979-987.

Vet A, Laven JSE, de Jong FH, Themmen APN, Fauser BCJM. Antimullerian hormone serum levels: a putative marker for ovarian aging. Fertil Steril 2002;77:357-62.

Vigier B, Tran D, Legeai L, Bezard J, Josso N. Origin of anti-Mullerian hormone in bovine freemartin fetuses. J Reprod Fertil 1984;70:473-9.

Webber LJ, Stubbs S, Stark J et al. Formation and early development of follicles in the polycystic ovary. Lancet 2003;362:1017-21.

Zaidi J, Barber J, Kyei Mensah A et al. Relationship of ovarian stromal blood flow at the baseline ultrasound scan to subsequent follicular response in an in vitro fertilization program. Obstet Gynecol 1996a;88:779-84.

Zaidi J, Collins W, Campbell S et al.Blood flow changes in the intraovarian arteries during the periovulatory period: relationship to the time of day. Ultrasound.Obstet.Gynecol 1996b;7:135-40.

Chapter 2

Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study

J. Kwee[1], M.W. Elting[1], R. Schats[1], P.D. Bezemer[2], C.B. Lambalk[1] and J. Schoemaker[1],

[1] Research Institute for Endocrinology, Reproduction and Metabolism. Department of Obstetrics and Gynaecology, division of Reproductive Endocrinology and Fertility and the IVF Centre, Vrije Universiteit Medical centre, Amsterdam, the Netherlands.

[2] Department of Clinical Epidemiology and Biostatistics Vrije Universiteit Medical centre, Amsterdam, the Netherlands.

Human Reproduction 2003;18:1422-7

Abstract

Background: This study was designed to compare endocrine tests (Clomiphene citrate Challenge Test (CCCT), Exogenous FSH Ovarian Reserve Test (EFORT) and basal FSH, basal E2, basal Inhibin B as an integral part of all CCCT's and EFORT's), with respect to their ability to estimate the stimulable cohort of follicles in the ovaries (ovarian capacity) and to analyse which test or combination of tests would give the best prediction of ovarian capacity.

Methods: One hundred and ten regularly menstruating patients, aged 18-39 years, participated in this prospective study, randomized, by a computer designed 4-blocks system study into two groups. Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by an IVF treatment. The result of ovarian hyperstimulation during IVF treatment, expressed by the total number of follicles, was used as golden standard.

Results: Univariate linear regression analysis showed that the best correlation with the number of follicles after ovarian hyperstimulation (Y) is found by the Inhibin B-increment in the EFORT ($Y = 3.957 + 0.081$ x InhB-incr. (95 % CI 0.061-0.101); $r = 0.751$; $P < 0.001$). Multiple linear regression analysis showed a significant contributing value of the variables bFSH, E2-increment of the EFORT and Inhibin B-increment to the basic model with the variable: age. The best prediction of ovarian capacity (Y) was seen, when E2-increment and Inhibin B-increment were used simultaneously in a stepforward multiple regression prediction model ($Y = 2.659 + 0.052$ x Inh B-incr.(0.026-0.078) + 0.027 x E2-incr.(95% CI 0.012-0.054); $r = 0.796$; $P < 0.001$). The CCCT could not be used in a prediction model.

Conclucions: The EFORT is the endocrine test which gives the best prediction of ovarian capacity.

Key Words: bFSH/bInhibin B/CCCT/EFORT/ovarian ageing/ovarian capacity

Introduction

Ageing of the ovary plays the major role in reproductive ageing and is related to the gradual reduction in the number of primordial follicles. The number of follicles leaving the pool of the so-called resting follicles to enter the growth phase towards the antral stages of development decreases with increasing age, leading to a stock at menopause estimated between less than 100 and 1000 primordial follicles in the pool (Gougeon *et al.*, 1994, Gougeon 1996). Scheffer *et al.* (Scheffer *et al.*, 1999) demonstrated that the number of primordial follicles in the ovary, as published by Faddy and Gosden. (Faddy and Gosden, 1996) correlated well with the number of growing follicles, counted by transvaginal sonography in the early follicular phase. So the decreasing size of the antral follicle cohort with age is a reflection of the decreasing primordial follicle pool. We can use this principle to measure ovarian capacity, defined as the total number of follicles which can be stimulated under maximal ovarian stimulation with FSH. A number of the so called ovarian capacity tests are supposed to indirectly reflect the size of the cohort of small antral follicles (2-5 mm in diameter) in the ovary. Van der Meer *et al.* (Van der *Meer et al.*, 1998) showed that in eumenorrheic patients, the median (range) FSH threshold level for monofollicular growth was 5.3 (4.3-8.2) IU/l and the median (range) threshold dose was 75 IU (0.5-1.75) FSH/day, given intravenously (i.v). It was concluded that by an increment of ½ ampoule of FSH (37.5 IU) above the threshold dose for monofollicular growth, the maximum response is already obtained. It seems that in IVF stimulation maximal effect is reached with FSH dosages up to 225 IE (The Latin-American puregon IVF Study Group, 2001, Out *et al.*, 2000, Out *et al.*, 2001). Combining these facts, it can be concluded that an initial stimulation by 3 ampoules of 75 IU of FSH under a long (GnRH agonist suppressed) protocol, probably gives a maximal IVF stimulation, the outcome of which could be used as the golden standard for the cohort size.

Endocrine tests predicting the ovarian capacity are either static: age (Hughes *et al.*, 1989, Meldrum, 1993, Navot *et al.*, 1994, Scott *et al.*, 1995), basal FSH (bFSH) (Pearlstone *et al.*, 1992, Toner, 1993, Cahill *et al.*, 1994, Hansen *et al.*, 1996), basal E2 (bE2) (Evers *et al.*, 1998, Smotrich *et al.*, 1995, Licciardi *et al.*, 1995), basal Inhibin B (bInhibin B) (Lahlou *et al.*, 1999)), or dynamic: Clomiphene citrate Challenge Test (CCCT) (Navot *et al.*, 1987, Loumaye *et al.*, 1990, Scott *et al.*, 1993), Exogenous FSH Ovarian Reserve Test (EFORT) (Fanchin *et al.*, 1994, Elting *et al.*, 2000)), GnRHa stimulation test (GAST) (Padilla *et al.*, 1990, Ravhon *et al.*, 2000). All tests predict the response to ovarian hyperstimulation and the prognosis for pregnancy in IVF treatment. Elting *et al.* (Elting *et al.*, 2000), showed that the EFORT could predict the follicle cohort size in patients with polycystic ovary syndrome, regularly menstruating women with polycystic ovaries and regularly menstruating women with normal ovaries. Except for the latter, none of the above tests have in fact been developed for determination of ovarian capacity.

The primary aim of this study was to compare endocrine tests with respect to their ability to measure the stimulative cohort of the ovaries (ovarian capacity). For reasons mentioned above the outcome of hyperstimulation with 3 ampoules in IVF under a long protocol was used as golden standard. The secondary aim of the study was to analyse which test or combination of tests would give the best prediction of ovarian capacity. For practical reasons the most direct stimulation of follicle growth (EFORT) was compared with the most indirect test (CCCT).

Materials and Methods

Study Population

One hundred and ten patients, aged 18-39 years, who were eligible for treatment by assisted reproduction between june 1997 to december 1999 participated in the study. This study is part of a prospective randomized study of regular menstruating patients to the determination of ovarian capacity called the DOC study. Their infertility was either idiopathic for > 3 years and/or due to a male factor and/or cervical hostility. Patients had to have regular menstrual cycles, two ovaries and at least one patent Fallopian tube. Excluded were patients with either polycystic ovary syndrome or a severe male factor, defined as 1. less than 1 million motile spermatozoa after Percoll centrifugation (gradient 40/90) and/or 2. > 20 % antibodies present on the spermatozoa after processing with Percoll centrifugation (gradient 40/90) and/or 3. > 50 % of the spermatozoa without an acrosome. Other exclusion criteria were untreated or insufficiently corrected endocrinopathies, clinically relevant systemic diseases or a body mass index > 28 kg/m².

The protocol was approved by the Committee on ethics of research involving human subjects of the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands. Informed consent was signed by all the couples participating in the study.

Treatment protocol

Patients were randomized by a computer designed 4-blocks system into two groups. Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by an IVF treatment under a long protocol. The bFSH level, bE2 level and bInhibin B level were determined as an integral part of all CCCT's and EFORT's.

Clomiphene citrate Challenge Test (CCCT): starting on the fifth day of the menstrual cycle (CD 1 = day of onset of menses) 100 mg of Clomiphene citrate (Serophene®; Ares Serono, Geneva, Switzerland) was administered for 5 days. In this study on CD 2 or 3 (basal values) and on CD 10 (stimulated values) the serum FSH, E2 and Inhibin B were determined. Analysis of the CCT included the following parameters: 1. bFSH and stimulated FSH (sFSH), 2. bE2 and stimulated E2 (sE2) and 3. bInhibin B and stimulated Inhibin B (sInhibin B).

Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT): on CD 3, 300 IU recFSH (Gonal-F®, Ares Serono, Geneva, Switzerland) were administered subcutaneously (s.c). In this study blood samples for the determination of FSH, E2 and Inhibin B were drawn: just before (basal values) and 24 hrs after (stimulated values) the administration of FSH. Analysis of the EFORT included the following parameters: 1. the bFSH, 2. E2-increment and Inhibin B-increment in 24 hrs after administration of FSH.

IVF-treatment: The ovarian hyperstimulation protocol was performed according to a long GnRH-agonist protocol starting in the midluteal phase. On CD 3 of the first cycle the CCT or the EFORT was performed as described above. In the subsequent midluteal phase, seven days after ovulation, daily s.c. injections with triptoreline-acetate (Decapeptyl®, 0.1 mg/day; Ferring, Hoofddorp, the Netherlands) were started. Because of the administration of the GnRH-agonist, patients were advised to use a barrier type of contraception during this cycle. On CD 3 of the next cycle, ovarian hyperstimulation was started with daily s.c. injections of a fixed dose of 225 IU uFSH (Metrodin HP®, 75 IU/amp; Ares Serono, Geneva, Switzerland),

because this dosage probably gives a maximal effect in follicle stimulation (Out *et al.*, 2000, Out *et al.*, 2001). Standard procedures were followed including transvaginal sonography (TVS) (Aloka SSD-1700, 5.0 MHz probe) on CD 2 or 3 and on CD 9 or 10. Daily TVS was performed from the moment when the leading follicle reached a diameter of 16 mm. Ovarian hyperstimulation was continued until the largest follicle reached a diameter of at least 18 mm. The maximum duration of uFSH administration was not allowed to exceed 16 days. If these criteria were met, Metrodin HP® and Decapeptyl® were discontinued and 10.000 IU of hCG (Profasi®, 10.000 IU/amp; Ares Serono, Geneva, Switzerland) were administered. On the day of hCG, TVS was performed to count the result of ovarian hyperstimulation (all follicles ≥ 10 mm) expressed as the total number of follicles.

Serum assay

Serum estradiol (E2) and FSH were determined by commercially available immunometric assays (Amerlite, Amersham, UK). For E2, the inter-assay CV was 11 % at 250 pmol/l and 8 % at 8000 pmol/l, the intra-assay coefficient of variation (CV) was 13 % at 350 pmol/l, 9 % at 1100 pmol/l and 9 % at 5000 pmol/l. The lower limit of detection for E2 was 90 pmol/l. In the EFORT and CCT we measured estradiol by a sensitive radioimmunoassay (Sorin, Biomedica, Saluggia, Italy). This measurement of estradiol was abbreviated as EE. For EE, the inter-assay CV was 11 % at 60 pmol/l, 8 % at 200 pmol/l, 11 % at 550 pmol/l and 8 % at 900 pmol/l. The intra-assay CV was 4 % at 110 pmol/l and 5 % at 1000 pmol/l. The lower limit of detection for EE was 18 pmol/l. For FSH, the inter-assay CV was 9 % at 3 IU/l and 5 % at 35 IU/l, the intra-assay CV was 9 % at 5 IU/l, 8 % at 15 IU/l and 6 % at 40 IU/l. The lower limit of detection for FSH was 0.5 IU/l. Inhibin B was determined immunometrically by a commercially available assay (Serotec Limited Oxford UK). For Inhibin B, the inter-assay CV was 17 % at 25 ng/L, 14 % at 55 ng/L and 9 % at 120 ng/L and the intra-assay CV was 8 % till 40 ng/l and 5 % at > 40 ng/l. The lower limit of detection for Inhibin B was 13 ng/l.Half way through the study, the Amerlite assay (suddenly withdrawn from the market) used to assess FSH had to be replaced by another commercially available assay (Delfia, Wallac, Finland). The two assays have been compared and showed excellent linear correlation, although a shift in the values took place. Delfia assay in comparison to Amerlite: Delfia FSH = 1.28 x Amerlite FSH + 0.01 (r=0.9964). For the Delfia FSH, the inter-assay CV was 5 % at 3.5 IU/l and 3 % at 15 IU/l. All FSH determinations have been recalculated and are expressed according to the Delfia assay. Values below the detection limit of an assay were assigned a value equal to the detection limit of that assay.

Statistical analysis

The endpoint of the study was the result of ovarian hyperstimulation expressed as the total number of follicles. Statistical analysis of all the data was performed with SPSS (Statistical Package for Social Sciences; SPSS, Inc., Chicago, IL) for Windows.

For the CCT-results, we used the variable or combination of variables showing the best correlation coefficient (Pearson's correlation test) with the total number of follicles obtained after stimulation. Univariate correlations between the variables: sFSH , sE2, sInhibin B, $\sum$ bFSH + sFSH , $\sum$ bE2 + sE2, $\sum$ bInhibin B + sInhibin B, FSH-increment in 7 days (sFSH-level - bFSH), E2-increment (sE2 - bE2) in 7 days, Inhibin B-increment (sInhibin B - bInhibin B) in 7 days versus the total number of follicles obtained after stimulation were analysed by Pearson's correlation test. Multivariate correlations between the above described variables

and the total number of follicles obtained after stimulation were analysed in a stepwise regression analysis.
For the EFORT-results, we analysed, if the bFSH had an additional contribution to the predictive value of the number of stimulated follicles already established by the E2-increment in 24 hrs or the Inhibin B-increment in 24 hrs, by stepwise linear regression analysis.
Comparison of means was done with the unpaired t-test or Wilcoxon's rank sum test.
By univariate linear regression, we estimated the value of the independent variables: age, bFSH, bE2, bInhibin B, CCT-results, E2-increment and the Inhibin B-increment in predicting the ovarian response.
We built a model based on the simplicity of the diagnostic tests at four different levels. Level 1: age, level 1-2: age and bFSH, level 1-2-3: age, bFSH and outcome of CCT or E2-increment in the EFORT. Level 1-2-3-4 (only for the EFORT-group): age, E2-increment in the EFORT and Inhibin B-increment in the EFORT. In a multiple regression model we estimated the additional value of the basal values (bFSH, bE2, binhibin B), the CCT and the EFORT on top of the basic model of age.
Stepwise regression analysis was used to find a prediction model for the ovarian response. The R square of the correlation of these variable(s) with the total number of follicles obtained after stimulation., reflects the proportion of the variability of the number of follicles explained by this variable(s). For all tests the significance level was 0.05.

Results

The characteristics of the 2 groups were depicted as means ± SD in Table I. No significant differences were noted between the groups in baseline characteristics, cycle day 3 measurements or outcome parameters. In the CCT group, 57.1% had primary infertility and 42.9% secondary infertility. The cause of infertility was 62.5% idiopathic, 28.6% male factor and 8.9% cervical factor. In the EFORT group, 65.0% had primary infertility and 35.0% secondary infertility. Their cause of infertility was 55.5% idiopathic, 42.5% male factor and 2% cervical factor.

Univariate linear regression analysis
The correlations between the CCT and number of follicles obtained after stimulation are calculated, the bFSH and the ∑ bFSH + sFSH show the best correlation coefficients ($r = 0.508$, $p < 0.001$ and $r = 0.496$, $p < 0.001$, respectively). In the further analysis we used the latter variable as the CCT-result. The regression line of the bFSH on the number of follicles was drawn by the regression equation: $Y = 30.334 – 2.114 \times bFSH$; with a 95% confidence interval of 1.135 - 3.092, meaning that each FSH increment of 1 IU/l predicts a decrement of 2.1 follicles (95 % CI: 1.1- 3.1). The regression equation for the ∑ bFSH + sFSH ,$Y = 25.626 - 0.712 \times CCT$ (0.372-1.052), shows that an increase of 1 IU/l predicts a decrement of 0.7 follicles. Also the correlation between age and the outcome parameter was highly significant ($Y = 56.500 – 1.250 \times age$ (0.631-1.869), $r = 0.482$, $p < 0.001$). The correlation between bInhibin B and the outcome parameter in the CCT-group was significant ($Y = 5.985 + 0.089 \times bInh\ B$ (0.024-0.153), $r = 0.351$, $p = 0.008$). bE2 and the endpoint were not significantly correlated ($Y = 14.360 - 0.0007 \times bE2$(-0.053-0.052), $r = 0.004$, $p = 0.978$).

Table I Characteristics of the groups (values are means ± SD). No significant differences.

	CCT-group N = 56	EFORT-group N = 54
Baseline characteristics		
Age (y)	33.79 ± 3.95	34.19 ± 3.75
Duration infertility (y)	3.71 ± 2.08	3.87 ± 1,56
Cycle day 3		
FSH (IU/l)	7.60 ± 2.46	7.38 ± 3.11
E2 (pmol/l)	126.05 ± 53.10	118.60 ± 47.06
Inhibin B (ng/l)	94.95 ± 39.36	96.33 ± 40.60
Treatment results		
Duration of stimulation (d)	12.4 ± 2.7	11.9 ± 2.3
Number of ampoules of FSH	34.2 ± 8.0	32.7 ± 7.0
E2 level on the day of hCG (pmol/l)	11155.41 ± 18591.13	12134.78 ± 17872.12
Endpoints		
Total number of follicles	14.27 ± 10.23	14.17 ± 10.27
Total number of oocytes	11.58 ± 8.51	11.93 ± 9.11

Figure 1A . Plot of the number of follicles obtained after stimulation against the E2-increment. The three lines represent the regression line: Y = 4.764 + 0.062 x E2-incr. with the 95 % confidence interval of the mean.

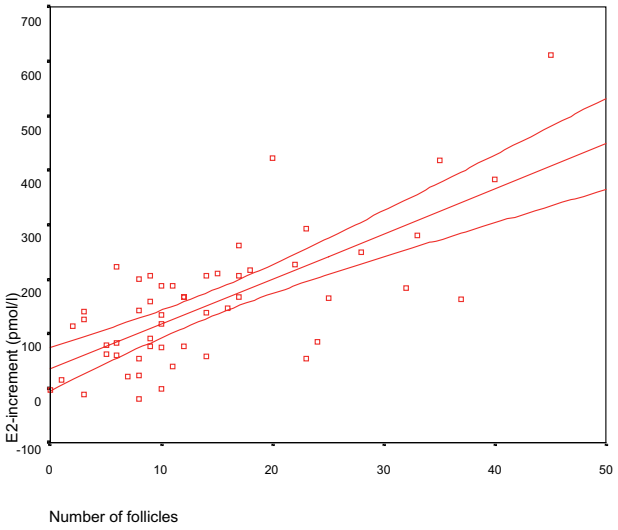
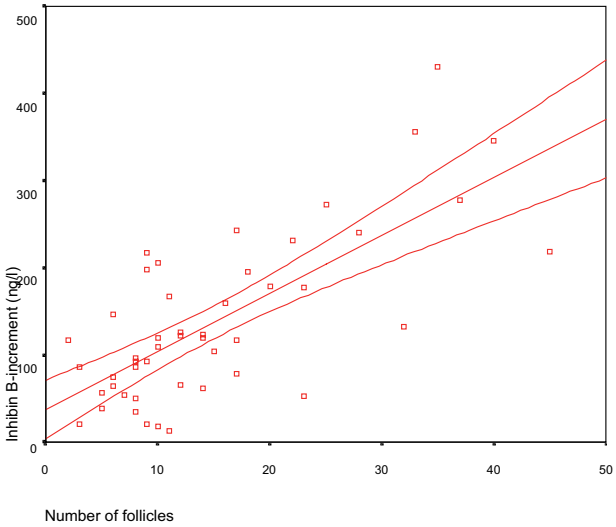


Figure 1B. Plot of the number of follicles obtained after stimulation against the Inhibin B-increment. The three lines represent the regression line: Y = 4.044 + 0.080 x InhB-incr. with the 95 % confidence interval of the mean.



In the EFORT group, the Inhibin B-increment and E2-increment in the EFOR-test show the best correlation coefficients (r = 0.751, p < 0.001 and r 0.718, p <0.001, respectively) with the total number of follicles obtained after stimulation. The regression line of the Inhibin B-increment on the number of follicles was drawn by the regression equation: Y = 3.957 + 0.081 x Inhibin B-increment; with a 95% confidence interval of 0.061 - 0.101, meaning that each Inhibin B-increment of 100 ng/l predicts 8.0 more follicles (95 % CI: 6.1-10.1) (figure 1A). The regression equation for the E2-increment (Y = 4.764 + 0.062 x E2-incr. (0.045-0.079)) shows that an increase of 100 pmol/l predicts 6.2 more follicles (figure 1B). Also the correlations between bFSH (Y = 17.374 – 0.370 x bFSH (0.063-0.677), r = 0.318) and age (Y = 48.597 – 1.004 x age (0.288-1.720), r = 0.364) with the outcome parameter were significant (p = 0.019, p = 0.007, respectively). The correlation between the bE2 (Y = 17.857 – 0.030 x bE2 (0.092–0.032), r = 0.134) and bInhibin B (Y = 9.055 + 0.059 x bInh B (0.001-0.119), r=0.266) with the endpoint were not significant (p = 0.340, p = 0.052, respectively).

Multiple linear regression analysis

For this analysis the order of parameters was established by the simplicity of the diagnostic tests: 1. age, 2. bFSH, 3. ∑ bFSH + sFSH for the CCT and 1. age, 2. bFSH, 3. E2-increment and 4. Inhibin B-increment for the EFORT. We excluded bE2 and bInhibin B, because bE2 did not correlate in either group and bInhibin B only correlated in the CCT-group. Table II shows the contributing value of each of the variables for the CCT-group. There was a significant contribution of bFSH to the model of age alone. Thereafter the ∑ bFSH + sFSH showed no further significant contribution to the model.

Table III shows the contributing value of each of the variables for the EFORT-group. There was a significant contribution of the bFSH to the basic model with the variable: age. When adding E2-increment and Inhibin B-increment as variables in multiple regression analysis to the basic model with age and bFSH, each of these showed a further significant contribution.

Table II Model based on the simplicity of the diagnostic tests for the CCT-group. 1 = Age, 2 = bFSH, 3 = ∑ bFSH + sFSH

Model	R	R square	Contributing P-value
1	0.482	0.233	< 0.001
1-2	0.615	0.378	0.001
1-2-3	0.629	0.396	0.217

Table III Model based on the simplicity of the diagnostic tests for the EFORT-group 1 = Age, 2 = bFSH, , 3 = E2-increment, 4 = Inhibin B-increment

Model	R	R square	Contributing P-value
1	0.364	0.132	0.007
1-2	0.445	0.193	0.046
1-2-3	0.748	0.559	< 0.001
1-2-3-4	0.806	0.649	0.001

Stepforward regression analysis: Prediction model for ovarian capacity

Based on the CCT group, the prediction model for ovarian response is explained for 25 % by the best predictive variable, the bFSH. When adding the independent variables: ∑ bFSH + sFSH and age in a stepforward regression analysis, the explained variation rose significantly with 12 % after the selection of age. The independent variable ∑ bFSH + sFSH did not have a significant contribution to the model. The exact prediction of the total number of follicles obtained after stimulation thus increased from 25 % to 37 %. The regression line of the bFSH and age on the number of follicles was drawn by the regression equation: Y = 58.139 – 1.644 x bFSH (0.684 -2.603) – 0.927 x age (0.323 -1.525) (r=0.605, p<0.001).

Based on the EFORT group, the prediction model for ovarian response is explained for 56 % by the best predictive variable, the Inhibin B-increment. When E2-increment and Inhibin B-increment were used simultaneously in a stepforward multiple regression prediction model, the explained variation of the best predictive model rose significantly with 7 %. The total explained variation thus increased from 56 % to 63 %. The regression line of the Inhibin B-increment and E2-increment on the number of follicles was drawn by the regression equation: Y = 2.659 + 0.052 x Inh B-incr.(0.026-0.078) + 0.027 x E2-incr.(0.012-0.054) (r=0.796, p<0.001). That means that if we use this formula, the confidence interval of Y is 50%. When we included age and bFSH as variables in the stepforward regression analysis together with the Inhibin B-increment and E2-increment we did not find a significant contribution of these variables.

Discussion

Our study shows that the Inhibin B-increment and E2-increment in the EFOR-test are the best predictors of the total number of follicles obtained after maximal ovarian hyperstimulation in an IVF-treatment i.e. cohort size. Age, bFSH, bE2, bInhibin B and the outcome of the CCT (∑ bFSH + sFSH) in this respect each, and in combination, show a much lower performance. This is in agreement with the study of Elting *et al*. (Elting *et al*., 2000) who concluded that the EFORT could predict the follicle cohort size in patients with the polycystic ovary syndrome, regularly menstruating women with polycystic ovaries and regularly menstruating women with normal ovaries. The best prediction model results when these 2 variables are used simultaneously in a stepforward multiple regression analysis. In 1994, Fanchin *et al*. (Fanchin *et al.*, 1994) described the EFORT, as a test which can detect a possible poor response among patients going to be treated with IVF. Our study showed that, in addition to finding poor responders, the test is able to predict with reasonable accuracy the number of follicles obtained after stimulation. This may, in combination with threshold analysis, be a first step to really control the number of follicles obtained after stimulation. As it has been suggested that granulosa cells of small antral follicles under the influence of FSH produce Inhibin B too, we were not surprised to find, as Elting *et al*. (Elting *et al.*, 2000) did, that the Inhibin B-increment follows the same pattern as the E2-increment in the EFOR-test.

Because we wanted to know if these tests had an additional value above age and the basic measurements, we also built a model for prediction of follicle number based on the simplicity of the diagnostic tests. The results show that there is a huge additional value for the E2-increment as well as for the Inhibin B-increment in the EFORT. There is no such additional value, however, for the best outcome parameter of the CCCT (∑ bFSH + sFSH). Navot *et al*. (Navot *et al.*, 1987) in 1987 described the use of the CCCT for the distinction between a poor and adequate response after ovarian hyperstimulation and its prognosis for pregnancy. It may well have its value there but for the prediction of the cohort size, the CCCT is of no use.

Our study also showed that, there was hardly any difference between the predictive value of age and bFSH for the number of follicles obtained after stimulation. Several studies show (Pearlstone *et al.*, 1992, Toner, 1993, Cahill *et al.*, 1994, Hansen *et al.*, 1996) that the bFSH as compared to a woman's age has a better predictive value of finding poor responders.

Unexpectedly we found no additional value of bInhibin B. This is in contrast with the finding of Seifer *et al.* (Seifer *et al.*, 1996, Danforth *et al.*, 1998, Seifer *et al.,*1999). They found that poor responders with normal bFSH levels have significantly lower bInhibin B levels than normal responders. Basal Inhibin B was also significantly correlated with chronological age, the number of ampoules of FSH administered, peak estradiol concentration, number of oocytes and embryos, and cancellation rates. We expect this difference to be caused by the fact that the Inhibin B production is strongly dependent on FSH (Hansen *et al.*, 1996). Inhibin B concentrations rise across the luteal-follicular transition and peak in the mid-follicular phase, but a few days later than similar changes in FSH, suggesting secretion by the granulosa cells of the developing cohort of follicles in response to FSH. Theoretically, under exogenous stimulation Inhibin B may be the optimal reflection of ovarian secretory capacity and follicle number. Therefore it could be that there is a better correlation during the mid-follicular phase, when granulosa cell function is strongly dependent on FSH compared with early follicular phase when FSH stimulation is strict marginal. This needs further investigation however.

A cost effectiveness analysis is currently under way. The average costs of an FSH stimulation

test will be more expensive than a CCCT. However it is not unlikely that due to better prediction of outcome, more accurate dose adjustment will reduce the overall costs due to limitations of gonadotropin use during stimulation and less cancelled cycles.
In conclusion, the results of our study show that in comparing endocrine tests for the prediction of the total number of follicles obtained after stimulation, Inhibin B-increment and E2-increment in the EFORT gave the best predictive values. Secondly the combination of Inhibin B-increment and E2-increment can predict ovarian capacity in regularly menstruating women, who are eligible for ART. The CCCT, measured in our study by the $\sum$ bFSH + sFSH, has no additional value above the basal values and age for the prediction of the number of follicles obtainable under maximal stimulation for IVF.

Acknowledgements
The authors acknowledge the help of Dr Corry Popp-Snijders and her staff, particularly for the endocrine laboratory work and the staff of the IVF centre for assistance during the execution of the protocol. This study was financially supported by Ares-Serono, Geneva, Switzerland.

References

Cahill DJ, Prosser CJ, Wardle PG, Ford WCL and Hull MGR (1994) Relative influence of serum follicle stimulating hormone, age and other factors on ovarian response to gonadotrophin stimulation. Br J Obstet Gynaecol 101, 999-1002.

Danforth DR, Arbogast LK, Mroueh J, Kim MH, Kennard EA, Seifer DB and Friedman CI (1998) Dimeric inhibin: a direct marker of ovarian aging. Fertil Steril 70, 119-123.

Elting MW, Kwee J, Schats R, Rekers-Mombarg LT and Schoemaker J (2000) The rise of Estradiol and Inhibin B after acute stimulation with follicle-stimulating hormone predict the follicle cohort size in women with polycystic ovary syndrome, regularly menstruating women with polycystic ovaries, and regularly menstruating women with normal ovaries. J Clin Endocrinol Metab 86, 1589-1595.

Evers JL, Slaats P, Land JA, Dumoulin JL and Dunselman GA (1998) Elevated levels of basal estradiol-17β predict poor response in patients with normal basal levels of follicle-stimulating hormone undergoing in vitro fertilization. Fertil Steril 69, 1010-1014.

Faddy MJ and Gosden RG (1996) A model conforming the decline in follicle numbers to the age of menopause in women. Hum Reprod 11, 1484-1486.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A and Frydman R (1994) Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Human Reprod 9, 1607-1611.

Gougeon A, Ecochard R and Thalabard JC (1994) Age-related changes of the population of human ovarian follicles: increase in the disappearance rate of non-growing and early-growing follicles in aging women. Biol Reprod 50, 653-663.

Gougeon A. (1996) Regulation of ovarian follicular development in primates: facts and hypotheses. Endocr Rev 17, 121-155.

Hansen LM, Batzer FR, Gutmann JN, Corson SL, Kelly MP and Gocial B (1996) Evaluating ovarian reserve: follicle stimulating hormone and oestradiol variability during cycle days 2-5. Hum Reprod 11, 486-489.

Hughes EG, King C and Wood EC (1989) A prospective study of prognostic factors in in vitro fertilization and embryo transfer. Fertil Steril 51, 838-844.

Lahlou N, Chabbert-Buffet N, Christin-Maitre S, Le Nestour E, Roger M and Bouchard P (1999) Main inhibitor of follicle stimulating hormone in the luteal-follicular transition: inhibin A, oestradiol, or inhibin B? Hum Reprod 14, 1190-1193.

Licciardi FL, Liu HC and Rosenwaks Z (1995) Day 3 estradiol serum concentrations as prognosticators of ovarian stimulation response and pregnancy outcome in patients undergoing in vitro fertilization. Fertil Steril 64, 991-994.

Loumaye E, Billion JM, Mine JM, Psalti I, Pensis M and Thomas K (1990) Prediction of individual response to controlled ovarian hyperstimulation by means of a clomiphene citrate challenge test. Fertil Steril 53, 295-301.

Meldrum DR (1993) Female reproductive aging-ovarian and uterine factors. Fertil Steril 59, 1-5.

Navot D, Rosenwaks Z and Margalioth EJ (1987) Prognostic assessment of female fecundity. Lancet 2, 645-647.

Navot D, Drews MR, Bergh PA, Guzman I, Karstaedt A, Scott RT, Garrisi GJ and Hofmann GE (1994) Age-related decline in female fertility is not due to dimished capacity of the uterus to sustain embryo implantation. Fertil Steril 61, 97-101.

Out HJ, Braat DM, Lintsen BME, Gurgan T, Bukulmez O, Gokmen O, Keles G, Caballero P, Gonzalez JM, Fabregues F et al. (2000) Increasing the daily dose of recombinant follicle stimulating hormone (Puregon®) does not compensate for the age-related decline in retrievable oocytes after ovarian stimulation. Hum Reprod 15, 29-35.

Out HJ, David I, Ron-El R, Friedler S, Shalev E, Geslevich J, Dor J, Shulman A, Ben Rafael Z, Fisch B et al. (2001) A randomized, double-blind clinical trial using fixed daily dose of 100 or 200 IU of recombinant FSH in ICSI cycles. Hum Reprod 16, 1104-1109.

Padilla SL, Bayati J and Garcia JE (1990) Prognostic value of the early serum estradiol response to leuprolide acetate in in vitro fertilization. Fertil Steril 53, 288-294.

Pearlstone AC, Fournet N, Gambone JC, Pang SC and Buyalos RP (1992) Ovulation induction in women age 40 and older: the importance of basal follicle-stimulating hormone level and chronological age. Fertil Steril 58, 674-679.

Ravhon A, Lavery R, Michael S, Donaldson M, Margara R, Trew G and Winston R (2000) Dynamic assays of inhibin B and oestradiol following buserelin acetate administration as predictors of ovarian response in IVF. Hum Reprod 15, 2297-2301.

Scheffer GJ, Broekmans FJM, Dorland M, Habbema JD, Looman CW and te Velde ER (1999) Antral follicle counts by transvaginal sonography are related to age in women with proven natural fertility. Fertil Steril 72, 845-851.

Scott RT, Illions EH, Kost ER, Dellinger C, Hofmann GE and Navot D (1993) Evaluation of the significance of the estradiol response during the clomiphene citrate challenge test. Fertil Steril 60, 242-246.

Scott RT, Opsahl MS, Leonardi MR, Neall GS, Illions EH and Navot D (1995) Life table analysis of pregnancy rates in a general infertility population relative to ovarian reserve and patient age. Hum Reprod 10, 1706-1710.

Seifer DB, Gardiner AC, Ferreira KA and Peluso JJ (1996) Apoptosis as a function of ovarian reserve in women undergoing in vitro fertilization. Fertil Steril 65, 593-598.

Seifer DB, Lambert-Messerlian G and Hogan JW (1997) Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome. Fertil Steril 67, 110-114.

Seifer DB, Scott RT, Bergh PA, Abrogast LK, Friedman CI, Mack CK and Danforth DR (1999) Women with declining ovarian reserve may demonstrate a decrease in day 3 serum inhibin B before a rise in day 3 follicle-stimulating hormone. Fertil Steril 72, 63-65.

Smotrich DB, Widra EA, Gindoff PR, Levy MJ, Hall JL and Stillman RJ (1995) Prognostic value of day 3 estradiol on in vitro fertilization outcome. Fertil Steril 64, 1136-1140.

The Latin-American Puregon IVF Study Group (2001) A double-blind clinical trial comparing a fixed daily dose of 150 and 250 IU of recombinant follicle-stimulating hormone in women undergoing in vitro fertilization. Fertil Steril 76, 950-956.

Toner JP (1993) The significance of elevated FSH for reproductive function. Bail Clin Obstet Gynaecol 7, 283-295.

Van der Meer M, Hompes PGA, de Boer JAM, Schats R and Schoemaker J (1998) Cohort size rather than follicle-stimulating hormone threshold level determines ovarian sensitivity in polycystic ovary syndrome. J Clin Endocrinol Metab 83, 423-426.

Chapter 3

The Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of low and hyperresponders to in vitro fertilization (IVF).

J. Kwee, R. Schats, J. McDonnell, J. Schoemaker, and C.B. Lambalk.

Division of Reproductive Endocrinology and Fertility and the IVF Centre, department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.

Fertility Sterility, 2006;85;1714-22

Abstract

Study Objective: This study was designed to compare the exogenous FSH ovarian reserve test (EFORT) versus the clomiphene citrate challenge test (CCCT), basal FSH and basal inhibin B, with respect to their ability to predict poor and/or hyper responders in an IVF population.
Design: Prospective randomized controlled trial
Setting: Fertility centre of an university hospital
Patients: 110 patients undergoing their first IVF cycle, randomized, by a computer designed 4-blocks system study into two groups.
Interventions: Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by an IVF treatment.
Main outcome measure(s): Ovarian response, expressed by the total number of retrieved oocytes.
Results(s): Univariate logistic regression showed that the best predictor for poor response is the CCCT (ROC-AUC = 0.87), with maximal accuracy of 0.89. Multiple logistic regression analysis did not produce a better model in terms of improving the prediction of poor response. For hyper response, univariate logistic regression showed that the best predictor is the inhibin B-increment in the EFORT (ROC-AUC = 0.92), but with a low maximal accuray of 0.78. Again, multiple logistic regression analysis did not produce a better model in terms of predicting hyper response.
Conclusion(s): Our study, the first which compares the CCCT with the EFORT for the prediction of poor and hyper responders, shows that the CCCT is superior for identification of low responders. EFORT (Inhibin B-increment) is superior for prediction of hyper response at cost of a high rate of false positives. Neither of the two tests seem adequate to act alone for identification of both poor and hyper responders.

Key Words: bFSH/bInhibin B/CCCT/EFORT/IVF/ovarian reserve

Introduction

The percentage of the general population seeking help for infertility is growing. One of the reasons for this is the fact that, especially in the western world, women are postponing their pregnancies because of their career. Ovarian reserve diminishes with increasing age and is related to fecundity (Hughes *et al.*, 1989, Kwee *et al.*, 2003, Meldrum, 1993, Navot *et al.*, 1994, Scott *et al.*, 1995).
Women, starting an infertility work-up will undergo extensive testing, and a large proportion of them will require expensive and invasive therapies, including assisted reproductive technologies.
An adequate dose of gonadotropins does not always result in an adequate ovarian response. In case of a poor response, very few or even no follicles develop, resulting in cancellation or, if the treatment cycle is completed, in a reduced conception rate.
On the other hand, an exaggerated ovarian response can be dangerous, leading to the development of many follicles. This situation may also lead to cancellation because of the risk of an ovarian hyperstimulation syndrome (OHSS), a condition which can be life threatening. Therefore, the identification of low or high responder patients before the start of an expensive, time consuming and often stressful IVF-treatment cycle is of importance. It would enable clinicians to adjust the ovarian stimulation protocol on an individual basis. This policy may lead to an optimal chance for every couple, and as such may reduce the number of cancelled ART-treatment cycles because of poor or exaggerated ovarian responses. This information however, is best used for counseling patients regarding their individual chances for pregnancy.
Several tests have been published giving a prediction for a poor or adequate response after hyperstimulation and/or a prognosis for pregnancy: basal FSH (bFSH) (Cahill *et al.*, 1994, Scott *et al.*, 1989, Toner, 1993), Clomiphene citrate Challenge test (CCCT) (Navot, 1987, Loumaye *et al.*, 1990, Scott *et al.*, 1993), Exogenous FSH Ovarian Reserve Test (EFORT) (Kwee *et al.*, 2003, Elting *et al.*, 2000, Fanchin *et al.*, 1994, Dzik *et al.*, 2000) and the GnRHa stimulation test (GAST) (Padilla *et al.*, 1990). No tests have been described which predict exaggerated high responses.
In a previous study (Kwee *et al.*, 2003) we published the results of the comparison of endocrine tests for the prediction of the total number of follicles obtained after stimulation. In that study, linear regression analysis indicated that Inhibin B-increment and E2-increment in the EFORT gave the best predictive values.
In this present study, based on the same data, we compared the most direct dynamic test, the exogenous FSH ovarian reserve test (EFORT) with the most indirect test, the clomiphene citrate challenge test (CCCT), basal FSH and basal inhibin B, with respect to their ability to predict poor and/or hyper responders in an IVF population by logistic regression analysis. The aim of the study was to find one single simple test, which could identify poor, normal and hyper responders.

Materials and Methods

Study Population

One hundred and ten patients, aged 18-39 years, who were eligible for treatment by assisted reproduction between june 1997 to december 1999 participated in the study. This study is part of a prospective randomized study of regular menstruating patients to the determination of ovarian reserve (Kwee *et al.*, 2003). Their infertility was either idiopathic for > 3 years and/or due to a male factor and/or cervical hostility. Patients had to have regular menstrual cycles, two ovaries and at least one patent Fallopian tube. Excluded were patients with either polycystic ovary syndrome, defined as a combination of oligo- or amenorrhoea and an increased luteinizing hormone (LH) concentration in the presence of a normal follicle stimulating hormone (FSH) or a severe male factor, defined as 1. less than 1 million motile spermatozoa after Percoll centrifugation (gradient 40/90) and/or 2. > 20 % antibodies present on the spermatozoa after processing with Percoll centrifugation (gradient 40/90) and/or 3. > 50 % of the spermatozoa without an acrosome. Other exclusion criteria were untreated or insufficiently corrected endocrinopathies, clinically relevant systemic diseases or a body mass index > 28 kg/m².

The protocol was approved by the Institutional review Board and the Committee on ethics of research involving human subjects of the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands. Informed consent was signed by all the couples participating in the study.

Treatment protocol

Patients were randomized by a computer designed 4-blocks system into two groups. Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by an IVF treatment under a long protocol. The reason that we chose a randomized design instead of a design in which all cases underwent both the CCCT and the EFORT was to avoid bias as a result of becoming pregnant after the first test.

It seems that in IVF stimulation the maximal effect is reached with FSH dosages up to 225 IU per day (Latin-American Puregon IVF study group, 2001, Out *et al.*, 2000, Out *et al.*, 2001). Using these results, we concluded that an initial stimulation by 3 ampoules of 75 IU of FSH under a long (GnRH-agonist suppressed) protocol, probably gives a maximal IVF stimulation, the outcome in terms of number of follicles and oocytes of which could be used as the golden standard for the cohort size. All patients underwent an transvaginal sonography on CD 3 to identify ovarian cysts. When there was an ovarian cyst of > 20 mm, the cycle was canceled.

Clomiphene citrate challenge test (CCCT): starting on the fifth day of the menstrual cycle (CD 1 = day of onset of menses) 100 mg of Clomiphene citrate (Serophene®; Serono, Geneva, Switzerland) was administered for 5 days. In this study on CD 2 or 3 (basal values) and on CD 10 (stimulated values) the serum FSH was determined. Analysis of the CCCT (2) was performed by the parameter: $\sum$ bFSH + sFSH.

Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT): on CD 3, 300 IU recFSH (Gonal-F®, Serono, Geneva, Switzerland) were administered subcutaneously (s.c). In this study blood samples for the determination of FSH, E2 and Inhibin B were drawn: just before (basal values) and 24 hrs after (stimulated values) the administration of FSH. Analysis

of the EFORT (2) included the following parameters: E2-increment and Inhibin B-increment 24 hrs after administration of FSH.
The bFSH and binhibin B level was determined as an integral part of all CCCT's and EFORT's.

IVF-treatment: The ovarian hyperstimulation protocol was performed according to a long GnRH-agonist protocol starting in the midluteal phase. On CD 3 of the first cycle the CCT or the EFORT was performed as described above. In the subsequent midluteal phase, seven days after ovulation, daily s.c. injections with triptoreline-acetate (Decapeptyl®, 0.1 mg/day; Ferring, Hoofddorp, the Netherlands) were started. Because of the administration of the GnRH-agonist, patients were advised to use a barrier type of contraception during this cycle. On CD 3 of the next cycle, ovarian hyperstimulation was started with daily s.c. injections of a fixed dose of 225 IU uFSH (Metrodin HP®, 75 IU/amp; Serono, Geneva, Switzerland), because this dosage probably gives a maximal effect in follicle stimulation. Standard procedures were followed including transvaginal sonography (TVS) (Aloka SSD-1700, 5.0 MHz probe) on CD 2 or 3 and on CD 9 or 10. Daily TVS was performed from the moment when the leading follicle reached a diameter of 16 mm. Ovarian hyperstimulation was continued until the largest follicle reached a diameter of at least 18 mm. The maximum duration of uFSH administration allowed was 16 days. If these criteria were met, Metrodin HP® and Decapeptyl® were discontinued and 10.000 IU of hCG (Profasi®, 10.000 IU/amp; Serono, Geneva, Switzerland) were administered. On the day of hCG, TVS was performed to count the result of ovarian hyperstimulation (all follicles ≥ 10 mm) expressed as the total number of follicles. TVS guided follicular aspiration (FA) was performed 36 hours after hCG administration. Follicular aspiration was done under systemic analgesia (7.5 mg diazepam orally and 50-100 mg pethidine hydrochloreide intramuscularly), and all follicles present were aspirated.

Serum assay
Serum estradiol (E2) and FSH were determined by commercially available immunometric assays (Amerlite, Amersham, UK). For E2, the inter-assay CV was 11 % at 250 pmol/l and 8 % at 8000 pmol/l, the intra-assay coefficient of variation (CV) was 13 % at 350 pmol/l, 9 % at 1100 pmol/l and 9 % at 5000 pmol/l. The lower limit of detection for E2 was 90 pmol/l. In the EFORT and CCT we measured estradiol by a sensitive radioimmunoassay (Sorin, Biomedica, Saluggia, Italy). This measurement of estradiol was abbreviated as EE. For EE, the inter-assay CV was 11 % at 60 pmol/l, 8 % at 200 pmol/l, 11 % at 550 pmol/l and 8 % at 900 pmol/l. The intra-assay CV was 4 % at 110 pmol/l and 5 % at 1000 pmol/l. The lower limit of detection for EE was 18 pmol/l. For FSH, the inter-assay CV was 9 % at 3 IU/l and 5 % at 35 IU/l, the intra-assay CV was 9 % at 5 IU/l, 8 % at 15 IU/l and 6 % at 40 IU/l. The lower limit of detection for FSH was 0.5 IU/l. Inhibin B was determined immunometrically by a commercially available assay (Serotec Limited Oxford UK). For Inhibin B, the inter-assay CV was 17 % at 25 ng/L, 14 % at 55 ng/L and 9 % at 120 ng/L and the intra-assay CV was 8 % till 40 ng/l and 5 % at > 40 ng/l. The lower limit of detection for Inhibin B was 13 ng/l.
Half way through the study (after 62 patients), the Amerlite assay used to assess FSH was suddenly withdrawn from the market and had to be replaced by another commercially available assay (Delfia, Wallac, Finland). The two assays have been compared and showed

excellent linear correlation, although a shift in the values took place (Delfia FSH = 1.28 x Amerlite FSH + 0.01 (r = 0.9964)). For the Delfia FSH, the inter-assay CV was 5 % at 3.5 IU/l and 3 % at 15 IU/l. All FSH determinations have been recalculated and are expressed according to the Delfia assay. Values below the detection limit of an assay were assigned a value equal to the detection limit of that assay.

Statistical analysis

The outcome measure of the study was the result of ovarian hyperstimulation expressed as the number of retrieved oocytes.

We defined a 'poor' ovarian response as less than 6 oocytes after ovarian hyperstimulation in an IVF treatment and a 'hyper' response as more than 20 oocytes after such an IVF treatment. The lower value was based on our experience that we have a 50-60 % chance of fertilisation in our laboratory. Consequently, in order to be able to transfer 2 or 3 qualitatively good embryos, at least 6 oocytes are required.

Moreover, defining poor responders at a slightly higher (7 oocytes) or lower (5 oocytes) cut off yield comparable proportions of poor responders and will produce similar predictive values (data not shown).

We defined a hyper response when there were > 20 oocytes. This was based on the knowledge that the pregnancy rates do not increase when > 20 oocytes are retrieved (18). Moreover, such cases have a significant risk of a severe OHSS.

Using univariate logistic regression analyses, we examined the value of the individual variables: age, bFSH, CCCT-results, E2-increment -and the Inhibin B-increment in the EFORT in predicting a poor and/or hyper reponse after ovarian hyperstimulation in IVF. Subsequently multivariate logistic regression analyses was used to develop prediction models for the ovarian response. The area under the receiver operating characteristic curve (ROC-AUC), was computed to assess the predictive accuracy of the logistic models.

To define a 'normal' and an 'abnormal' test, sensitivity, specificity, positive predictive value and accuracy were used to find the optimal cut off level.

Comparison of means was done with the unpaired t-test. For all tests the significance level was 0.05.

Statistical analysis of the data was performed with SPSS (Statistical package for Social Sciences; SPSS, Inc., Chicago, IL) for Windows.

Results

The characteristics of the 2 groups are given as means ± SD in Table I. No significant differences were noted between the groups in baseline characteristics, cycle day 3 measurements or outcome parameters.

In the CCCT group, 57.1 % had a primary infertility and 42.9 % a secondary infertility. The cause of infertility was for 62.5 % an idiopathic factor, 28.6 % a male factor and 8.9 % a cervical factor.

In the EFORT group, 65.0 % had a primary infertility and 35.0% a secondary infertility. The cause of infertility was for 55.5 % an idiopathic factor, 42.5 % a male factor and 2 % a cervical factor.

In the CCCT group, 32 patients had a normal response to ovarian stimulation, 15 patients had a poor response and 9 patients had a hyper response. In the EFORT group 32 patients had a normal response to ovarian stimulation, 14 patients had a poor response and 8 patients had a hyper response.

Table I Characteristics of the groups (values are means ± SD). No significant differences (Kwee *et al*, 2003).

	CCT-group N = 56	EFORT-group N = 54
Baseline characteristics		
Age (y)	33.9 ± 4.0	34.2 ± 3.8
Duration infertility (y)	3.7 ± 2.1	3.9 ± 1.6
Cycle day 3		
FSH (IU/l)	7.6 ± 2.5	7.4 ± 3.1
E2 (pmol/l)	126.1 ± 53.1	118.6 ± 47.1
Inhibin B (ng/l)	95.0 ± 39.4	96.3 ± 40.6
Treatment results		
Duration of stimulation (d)	12.4 ± 2.7	11.9 ± 2.3
Number of ampoules of FSH	34.2 ± 8.0	32.7 ± 7.0
E2 level on the day of hCG (pmol/l)	11155.4 ± 18591.1	12134.8 ± 17872.1
Endpoints		
Total number of follicles	14.3 ± 10.2	14.2 ± 10.3
Total number of oocytes	11.6 ± 8.5	11.9 ± 9.1

Table 2 depicts the statistical significance and areas under the receiver operating characteristic curve (ROC-AUC) of logistic regression analysis for 5 ovarian reserve tests for the prediction of poor response after IVF with ovarian hyperstimulation. As a single prognostic predictor, the CCCT appeared to have the best discriminative potential for poor response, as expressed by the largest ROC-AUC (0.88). The E2-increment in the EFORT had a ROC-AUC of 0.75. The Inhibin B-increment in the EFORT had a ROC-AUC of 0.86; this was not statistically different from the 0.88 of the CCCT ROC-AUC and the bFSH had a ROC-AUC of 0.83.
In the CCCT group, multivariate analysis resulted in a predictive logistic model with only one variable, namely the CCCT (ROC-AUC = 0.88). In the EFORT group, multivariate analysis resulted in a predictive logistic model with two variables, bFSH and the Inhibin B-increment in the EFORT (ROC-AUC = 0.87).

Table 2 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC-AUC) of the ovarian reserve tests for the prediction of 'poor' response in IVF.

Variable	N	P	ROC AUC
Univariate analysis			
Age (y)	110	0.033	0.63
bFSH (IU/l)	110	< 0.0001	0.83
bInhibin B (ng/l)	110	0.32	0.56
CCT (IU/l)	56	< 0.0001	0.88
E2-increment in the EFORT (pmol/l)	54	0.006	0.75
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.86
Multivariate analysis			
CCCT GROUP			
Age (y)	56	0.99	} 0.88
bFSH (IU/l)	56	0.70	
CCT (IU/l)	56	< 0.0001	
Multivariate analysis			
EFORT GROUP			
Age (y)	54	0.51	} 0.87
E2-increment in the EFORT (pmol/l)	54	0.42	
bFSH (IU/l) and Inh.B-incr.in the EFORT (ng/l)	54	< 0.0001	

NS = not significant

Table 3 presents, the tests characteristics sensitivity, specificity, positive predictive value and accuracy at different cut off levels of the CCCT, EFORT and bFSH used define a normal (non-poor response) and an abnormal (poor response) test for the prediction of 'poor' response after IVF. The cut off level of > 18 IU/l in the CCCT gave the highest value of the sum of sensitivity plus specificity and also gave the highest accuracy for the prediction of 'poor' response. This value had a sensitivity of 73 % and a specificity of 95 % and in the population studied, with a prevalence of 27 % for a poor response (< 6 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 89 % (which means that 89 % had a correctly predicted test). In case of a result greater than 18 IU/l in the CCCT, the test correctly predicted poor response to stimulation in an IVF-treatment in 85 % (positive predictive value).
As shown in table 3, all test characteristics of the E2 –and Inhibin B increment in the EFORT and bFSH at different cut off levels were of less clinical relevance comparing to the test characteristics of the CCCT at the cut off level of 18 IU/l.

Table 3 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for the CCCT (IU/l), E2-increment in EFORT (mmol/l), Inhibin B-increment in EFORT (ng/l) and bFSH (IU/l) for the prediction of 'poor' response in IVF.

CCCT(IU/l)	Sensitivity	Specificity	PPV	Accuracy
> 14	0.93	0.68	0.51	0.68
> 16	0.80	0.83	0.63	0.82
> 18	0.73	0.95	0.85	0.89
> 20	0.60	0.98	0.90	0.87
> 22	0.53	0.98	0.89	0.76

E2-increment in EFORT (mmol/l)	Sensitivity	Specificity	PPV	Accuracy
< 110	0.64	0.68	0.41	0.67
< 120	0.64	0.65	0.39	0.65
< 130	0.71	0.65	0.42	0.67
< 140	0.79	0.60	0.41	0.65
< 150	0.86	0.57	0.41	0.65

Inhibin B-increment in EFORT (ng/l)	Sensitivity	Specificity	PPV	Accuracy
< 70	0.64	0.80	0.53	0.76
< 80	0.71	0.78	0.53	0.76
< 90	0.79	0.75	0.52	0.76
< 100	0.86	0.70	0.50	0.74
< 110	0.93	0.68	0.50	0.72

bFSH (IU/l)	Sensitivity	Specificity	PPV	Accuracy
> 4	1.00	0.05	0.18	0.30
> 6	0.93	0.40	0.36	0.54
> 8	0.72	0.78	0.54	0.76
> 10	0.35	0.96	0.77	0.80
> 12	0.24	1.00	1.00	0.80

Table 4 depicts statistical significance and areas under the receiver operating characteristic curve of logistic regression analysis for the ovarian reserve tests for the prediction of hyper response after IVF with ovarian hyperstimulation. As a single prognostic predictor, the Inhibin-B increment in the EFORT appeared to have the best discriminative potential for hyper response, (ROC-AUC = 0.92).
By comparison the E2-increment in the EFORT had a ROC-AUC of 0.83, the CCCT had a ROC-AUC of 0.82 and the bFSH had a ROC-AUC of 0.80.
In the CCT group, multivariate analysis resulted in a model with two variables, age and the CCCT (ROC-AUC = 0.90). In the EFORT group, multivariate analysis resulted in a model with only one variable, the Inhibin B-increment in the EFORT (ROC-AUC = 0.92).

Table 4 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC-AUC) of the ovarian reserve tests for the prediction of 'hyper 'response' in IVF.

Variable	N	P	ROC AUC
Univariate analysis			
Age (y)	110	0.004	0.71
bFSH (IU/l)	110	< 0.0001	0.80
BInhibin B (ng/l)	110	0.052	0.65
CCT (IU/l)	56	0.003	0.82
E2-increment in the EFORT (pmol/l)	54	0.003	0.83
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.92
Multivariate analysis			
CCCT GROUP			
bFSH (IU/l)	56	0.46	} 0.90
bInhibin B (ng/l)	56	0.09	
Age (y) and CCT (IU/l)	56	< 0.0001	
Multivariate analysis			
EFORT GROUP			
Age (y)	54	0.65	} 0.92
bFSH (IU/l)	54	0.97	
bInhibin B (ng/l)	54	0.07	
E2-increment in the EFORT (pmol/l)	54	0.26	
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	

NS = not significant

Table 5 presents the same tests characteristics at different cut off levels of the CCCT, EFORT and bFSH to define a normal (non-hyper response) and an abnormal (hyper response) test for the prediction of hyper response after IVF. The cut off level of > 130 ng/l gave the highest sum of the sensitivity plus specificity and gave also the highest accuracy. This result had a

sensitivity of 100 % and a specificity of 74%. In the population studied, with a prevalence of 15 % for high response (> 20 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 78 % (which means that 78 % had a correctly predicted test). In case of a result greater than 130 ng/l for the Inhibin B-increment in the EFORT, the test correctly predicted hyper response to stimulation in an IVF-treatment in 40 % (positive predictive value).
As shown in table 5, all test characteristics of the E2-increment in the EFORT, CCCT and bFSH at different cut off levels were of less clinical relevance comparing to the test characteristics of the Inhibin B-increment in the EFORT at the cut off level of 130 ng/l.

Table 5 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for the CCCT (IU/l), E2-increment in EFORT (mmol/l), Inhibin B-increment in EFORT (ng/l) and bFSH (IU/l) for the prediction of 'hyper' response in IVF.

CCCT(IU/l)	Sensitivity	Specificity	PPV	Accuracy
< 9	0.23	0.98	0.67	0.86
< 10	0.44	0.96	0.67	0.88
< 11	0.44	0.89	0.44	0.82
< 12	0.67	0.75	0.33	0.73
< 13	078	0.62	0.28	0.64

E2-increment in EFORT (ng/l)	Sensitivity	Specificity	PPV	Accuracy
> 200	0.63	0.78	0.33	0.76
> 220	0.63	0.89	0.50	0.85
> 230	0.63	0.91	0.56	0.87
> 240	0.63	0.94	0.63	0.89
> 260	0.50	0.94	0.57	0.87

Inhibin B-increment in EFORT (ng/l)	Sensitivity	Specificity	PPV	Accuracy
> 120	1.00	0.37	0.33	0.70
> 125	1.00	0.72	0.38	0.76
> 130	1.00	0.74	0.40	0.78
> 135	0.88	0.74	0.37	0.76
> 140	0.88	0.74	0.37	0.76

bFSH (IU/l)	Sensitivity	Specificity	PPV	Accuracy
< 4	0.18	0.99	0.75	0.86
< 5	0.29	0.94	0.45	0.82
< 6	0.65	0.76	0.48	0.76
< 7	0.82	0.61	0.28	0.65
< 8	0.94	0.41	0.23	0.49

Discussion

Our study, the first which compares the CCCT with the EFORT for the prediction of poor and hyper responders, shows that although the EFORT was the best predictor for the total number of follicles after stimulation for IVF (Kwee *et al.*, 2003), it was not better in predicting poor responders than the CCCT. On the contrary, the test characteristics of the CCCT were much better to predict poor responders in an IVF population. On the other hand, the Inhibin B-increment in the EFORT was superior in identifying hyper responders.
We aimed to search for a single test for ovarian reserve, able to optimally identify all patients with an adequate response, a poor response and those with a very high response.
The consequence of finding a potentially poor responder in a group of patients with normal regular cycles, could be that we would start with gonadotrophin stimulation with a higher dose and we would not deny entry into the IVF program. It has been shown sufficiently that women with a diminished ovarian reserve may still become pregnant (Van Rooij *et al.*, 2004).
What would be the clinical consequences of false positive and false negative test results for the poor responder? False positive patients (those erroneously designated as poor responders), would be treated with a higher dose and are consequently exposed to the risk of exaggerated ovarian response and thus have an increased risk of developing OHSS. On the other hand, false negative patients would receive a suboptimal stimulation scheme. Both groups of patients are at risk for cycle cancellation. One could argue that the consequence related to a false positive result (OHSS) is more serious and threatening than the consequence of a false negative test result. For a test which attempts to identify poor responders, this implies that the specificity should be high, which means a high threshold and consequently a lower sensitivity and a higher number of false negative patients (Scott, 2004). Unfortunately the combination of an extremely high cut-off level, a poor sensitivity and a high specificity will make the number of patients who might benefit from the test very small.
Our comparative study showed that the CCCT (ROC-AUC = 0.88), Inhibin B-increment in the EFORT (ROC-AUC = 0.86) and the model with bFSH and Inhibin-B increment in the EFORT (ROC-AUC = 0.87) were the optimal prognostic tests to predict a poor response in IVF. At the cut off level of 18 IU/l in the CCCT, the accuracy was optimal (89 %) with a sensitivity of 73 % and a high specificity of 95 %. This result was much higher than any cut off level in the Inhibin B-increment in the EFORT. We therefore conclude that the CCCT (cut off level of 18 IU/l) was the best test to predict a poor response in IVF.
It has to be mentioned that these values are lower than those reported by Bancsi et al. (2002) in logistic models for predicting low response combining antral follicle count, bFSH and bInhibin B (ROC-AUC = 0.92) suggesting that the dynamic tests might not be superior in this respect.
In patients with a potential hyper response, the clinical consequence would be to start with a lower dose of gonadotrophins to avoid OHSS (Hendriks *et al.*, 2004) since there is evidence that the ovarian response can be reduced by minimal stimulation protocols (Fauser *et al.*, 1999, Popovic-Todorovic *et al.*, 2003). False negative patients will run the risk of exaggerated ovarian response and thus have a risk of OHSS, while a false positive result will lead to a suboptimal stimulation scheme. Again both situations would have an increased risk of cycle cancellation, but the impact of having many false negative patients (with an increased risk of OHSS) will be reduced. Therefore, for this test, the sensitivity should be high, implying low cut off levels, a lower specificity and consequently a higher number of unrecognised non

high responding patients.
Our comparison showed that the Inhibin B-increment in the EFORT was the optimal prognostic test to predict a hyper response in IVF, with a high sensitivity at the cut off level of 130 ng/l. At that point the accuracy was the highest (78 %). Thus we conclude that the Inhibin B-increment (cut off level of 130 ng/l) seems the best test to predict a 'hyper' response in IVF, but at the cost of a high number of false positive patients.
Recently Hendriks *et al* (2004) studied the usefulness of measuring stimulated serum estradiol levels in predicting ovarian hyperresponse and he concluded that its clinical value is low.
In an earlier study (Kwee *et al.*, 2003) we found that the EFORT was superior to CCCT in predicting the outcome of maximal ovarian hyperstimulation, but this was over the whole range from low to high response without specified thresholds. The CCCT is inferior under such "whole range" conditions (Kwee *et al.*, 2003) and it has superior performance when a below threshold outcome needs to be identified as we show in the current study. This only means that a low to normal response of FSH on clomiphene does not always implicates a hyper response. This FSH level is the final reflection of the integral reproductive endocrine axis and thus rather indirectly a function of quantitative ovarian funtioning in particular when lower normal FSH levels are involved. The EFORT with its granulosa cell derived E2 and Inhibin B is a much more direct test for the number of growing antral follicles and thus able to cover the whole range.
Based on our earlier (Kwee *et al.*, 2003) and the current study, it would seem that the EFORT meets the criteria to be a test to predict poor and hyper responders. But in our view actually neither the CCCT nor the EFORT were of sufficient quality to act alone for identification of poor and hyper responders.
The CCCT is superior with regard to identification of potential low responders and EFORT is superior to identify hyper responders, but with a very high rate of false positive results. We therefore prefer the CCCT over EFORT as the most useful single test to perform for determination of ovarian response.
Future studies will have to be carried out to determine if other ovarian reserve tests such as GnRH agonist stimulation test (Padilla *et al.*, 1990, Ravhon *et al.*, 2000), the measurement of Anti- Müllerian Hormone (AMH) (Vet *et al.*, 2002, Hazout *et al.*, 2004, Seifer *et al.*, 2002) and antral follicle count (Scheffer *et al.*, 2003) are able to serve both purposes.

Acknowledgements
The authors acknowledge the help of Dr Corry Popp-Snijders and her staff, particularly for the endocrine laboratory work and the staff of the IVF centre for assistance during the execution of the protocol. This study was financially supported by Ares-Serono, Geneva, Switzerland.

References

Bancsi LF, Broekmans FJ, Eijkemans MJ, de Jong FH, Habbema JD, te Velde, E.R. Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steri l 2002;77:328-36.

Cahill DJ, Prosser CJ, Wardle PG, Ford WCL, Hull MGR. Relative influence of serum follicle stimulating hormone, age and other factors on ovarian response to gonadotrophin stimulation. Br J Obstet Gynaecol 1994;101: 999-1002.

Dzik A, Lambert-Messerlian G, Izzo VM, Soares JB, Pinotti JA, Seifer DB. Inhibin B response to EFORT is associated with the outcome of oocyte retrieval in the subsequent in vitro fertilization cycle. Fertil Steril 2000;74:1114-7.

Elting MW, Kwee J, Schats R, Rekers-Mombarg LT, Schoemaker J. The rise of Estradiol and Inhibin B after acute stimulation with follicle-stimulating hormone predict the follicle cohort size in women with polycystic ovary syndrome, regularly menstruating women with polycystic ovaries, and regularly menstruating women with normal ovaries. J Clin Endocrinol Metab 2000;86:1589-95.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A, Frydman R. Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Human Reprod 1994;9:1607-11.

Fauser BC, Devroey P, Yen SS, Gosden R, Crowley WF Jr, Baird DT, Bouchard P. Minimal ovarian stimulation for IVF: appraisal of potential benefits and drawbacks. Hum Reprod 1999;14:2681-6.

Hazout A, Bouchard P, Seifer DB, Aussage P, Junca AM, Cohen-Bacrie P. Serum antimullerian hormone/ mullerian-inhibiting substance appears to be a more discriminatory marker of assisted reproductive technology outcome than follicle-stimulating hormone, inhibin B, or estradiol. Fertil Steril 2004;82:1323-9.

Hendriks DJ, Klinkert ER, Bancsi LF, Looman CW, Habbema JD, te Velde ER, Broekmans FJ. Use of stimulated serum estradiol measurements for the prediction of hyperresponse to ovarian stimulation in in vitro fertilization(IVF). J Assist reprod Genet 2004;21:65-72

Hughes EG, King C, Wood EC. A prospective study of prognostic factors in in vitro fertilization and embryo transfer. Fertil Steril 1989;51:838-44.

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB, Schoemaker, J. Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVFtreatment: results of a prospective randomized study. Hum Reprod 2003;18:1422-7.

Latin-American Puregon IVF Study Group. A double-blind clinical trial comparing a fixed daily dose of 150 and 250 IU of recombinant follicle-stimulating hormone in women undergoing in vitro fertilization. Fertil Steril 2001;76: 950-6.

Loumaye E, Billion JM, Mine JM, Psalti I, Pensis M, Thomas K. Prediction of individual response to controlled ovarian hyperstimulation by means of a clomiphene citrate challenge test. Fertil Steril 1990;53:295-301.

Meldrum DR. Female reproductive aging-ovarian and uterine factors. Fertil Steril 1993;59: 1-5.

Navot D, Rosenwaks Z, Margalioth EJ. Prognostic assessment of female fecundity. Lancet 1987;2: 645-7.

Navot D, Drews MR, Bergh PA, Guzman I, Karstaedt A, Scott RT et al. Age-related decline in female fertility is not due to dimished capacity of the uterus to sustain embryo implantation. Fertil Steril 1994;61:97-101.

Out HJ, Braat DM, Lintsen BME, Gurgan T, Bukulmez O, Gokmen O et al. Increasing the daily dose of recombinant follicle stimulating hormone (Puregon®) does not compensate for the age-related decline in retrievable oocytes after ovarian stimulation. Hum Reprod 2000;15:29-35.

Out HJ, David I, Ron-El R, Friedler S, Shalev E, Geslevich J et al. A randomized, double-blind clinical trial using fixed daily dose of 100 or 200 IU of recombinant FSH in ICSI cycles. Hum Reprod 2001;16:1104-9.

Padilla SL, Bayati J, Garcia JE. Prognostic value of the early serum estradiol response to leuprolide acetate in in vitro fertilization. Fertil Steril 1990;53:288-94.

Popovic-Todorovic B, Loft A, Bredkjaeer HE, Nielsen IK, Andersen AN.. A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. Hum Reprod 2003;18:2275-82.

Ravhon A, Lavery S, Michael S, Donaldson M, Margara R, Trew G et al. Dynamic assays of inhibin B and oestradiol following buserelin acetate administration as predictors of ovarian response in IVF. Hum Reprod 2000;15:2297–2301.

Scheffer GJ, Broekmans FJM, Looman CWN, Blankenstein M, Fauser BCJM, de Jong FH, te Velde ER. The number of antral follicles in normal women with proven fertility is the best reflection of reproductive age. Hum Reprod 2003;18:700-6.

Scott RT, Toner JP, Muasher SJ, Oehninger S, Robinson S, Rosenwaks Z. Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. Fertil. Steril 1989;51:651-4.

Scott RT, Illions EH, Kost ER, Dellinger C, Hofmann GE, Navot D. Evaluation of the significance of the estradiol response during the clomiphene citrate challenge test. Fertil Steril 1993;60:242-6.

Scott RT, Opsahl MS, Leonardi MR, Neall GS, Illions EH, Navot D. Life table analysis of pregnancy rates in a general infertility population relative to ovarian reserve and patient age. Hum. Reprod 1995;10:1706-10.

Scott RT. Jr. Diminished ovarian reserve and access to care. Fertil Steril 2004;81:1489-92.

Seifer DB, Mac Laughlin DT, Christian BP, Feng B, Shelden RM. Early follicular serum mullerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. Fertil Steril 2002;77:468-71.

Toner JP. The significance of elevated FSH for reproductive function. Bail Clin Obstet Gynaecol 1993;7 : 283-95.

Van Rooij IA, de Jong E, Broekmans FJ, Looman CW, Habbema JD, te Velde ER. High follicle-stimulating hormone levels should not necessarily lead to the exclusion of subfertile patients from treatment. Fertil Steril 2004; 81:1478-95.

Vet A, Laven JSE, de Jong FH, Themmen APN, Fauser BCJM. Antimüllerian hormone serum levels: a putative marker for ovarian aging. Fertil Steril 2002;77:357-62.

Chapter 4

Intercycle variability of ovarian capacity tests: results of a prospective randomized study

J. Kwee, R. Schats, J. McDonnell, C.B. Lambalk and J. Schoemaker.

Division of Reproductive Endocrinology and Fertility and the IVF Centre, department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.

Human Reproduction 2004;19:590-5

Abstract

Background: This study was designed to prospectively assess the intercycle variability (ICV) of basal FSH (bFSH), Clomiphene citrate Challenge Test (CCT) (Analysis of the CCT was performed by the parameter: ∑ bFSH + sFSH) and Exogenous FSH Ovarian Reserve Test (EFORT) (Analysis of the EFORT included the following parameters: E2-increment and Inhibin B-increment 24 hrs after administration of FSH), and secondarily to assess the influence of the variability of these ovarian reserve tests.
Methods: Eighty five regularly menstruating patients, aged 18-39 years, participated in this prospective study, randomized, by a computer designed 4-blocks system into two groups. Fourthy three patients underwent a CCT, and 42 patients underwent an EFORT. Each test was performed 1-4 times in subsequent cycles, one test per cycle. During the first 3 cycles patients were treated with intra uterine inseminations (IUI). Follicle number and oocyte yield during IVF hyperstimulation in the 4th cycle were taken as measure for ovarian capacity.
Results: The per-cycle variance of bFSH ranged from 1.8 to 4.4 (maximum to minimum ratio of 2.44, $p<0.0001$), while that of CCT ranged from 21.3 to 70.6 (3.31, $p<0.0001$). No significant change in per cycle variance was found for the E2-increment (1.25, $p>0.2$) and Inhibin B-increment (1.31, $p>0.2$), which were the EFORT parameters. A large intercycle variability of CCT and basal FSH test results was strongly associated with lower ovarian capacity.
Conclusions: Our study shows that the intercycle variability of the Inhibin-B increment and the estradiol increment in the EFORT is stable in consecutive cycles, which indicates that this reproducible test is a more reliable tool for determination of ovarian reserve than bFSH and CCT. Women with limited ovarian reserve show a strong intercycle variability of bFSH and FSH response to clomiphene citrate.

Key Words: bFSH/CCT/EFORT/intercycle variability/ovarian ageing/ovarian reserve

Introduction

Basal FSH (bFSH) (Scott *et al.*, 1989), clomiphene citrate challenge tests (CCT) (Navot *et al.*, 1987, and the exogenous FSH ovarian reserve test (EFORT) (Fanchin *et al.*, 1994, Kwee *et al.*, 2003) have been shown to be of predictive value for the ovarian reserve with respect to hyperstimulation and pregnancy rates in patients undergoing in vitro fertilization (IVF) (Sharara *et al.*, 1998). Kwee *et al.* (Kwee *et al.*, 2003) compared the predictive value of bFSH, CCT and the EFORT on the outcome of ovarian hyperstimulation in IVF treatment and concluded that the EFORT was the best endocrine test for the prediction of ovarian reserve. A small number of studies have documented the intercycle variation of bFSH and CCT in the same patient. Scott et al. (Scott *et al.*, 1990) documented the intercycle variability of FSH and found it to vary from patient to patient. Hannoun et al. (Hannoun *et al.*, 1998) documented a high degree of intercycle variability of the CCT when performed in the same patient.
The knowledge of the intercycle variability of an ovarian reserve test is important for correct interpretation of test results. No data at all are available about the intercycle variability of the EFORT. The purpose of our study was to prospectively assess the intercycle variability of bFSH, CCT and EFORT, and secondarily to assess the influence of the variability of these tests on the prediction of ovarian reserve (ovarian capacity), defined by us as the maximal number of follicles which can be stimulated under maximal ovarian hyperstimulation with FSH.

Materials and Methods

Study Population

This study is part of a prospective randomized study of regularly menstruating patients on the determination of ovarian capacity, called the DOC study. From june 1997 till may 1999, 85 patients aged 18-39 years who were eligible for intra-uterine insemination (IUI), entered the study. Their infertility was either idiopathic for > 3 years and/or due to a male factor and/or cervical hostility. Patients had to have regular menstrual cycles, two ovaries and at least one patent Fallopian tube. Excluded were patients with either polycystic ovary syndrome or a severe male factor, subsequently treated by ICSI and defined as 1. < than 1 million motile spermatozoa after Percoll centrifugation (gradient 40/90) and/or 2. > 20 % antibodies present on the spermatozoa after processing with Percoll centrifugation (gradient 40/90) and/or 3. > 50 % of the spermatozoa without an acrosome. Other exclusion criteria were untreated or insufficiently corrected endocrinopathies, clinically relevant systemic diseases or a body mass index > 28 kg/m².
The study protocol was approved by the Committee on ethics of research involving human subjects of the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands. Informed consent was signed by all the couples participating in the study.

Treatment protocol

Patients were randomized by a computer designed 4-blocks system into two groups of 43 patients, for the study of the CCT, and 42 patients for the study of the EFORT. Each test was performed once per cycle for up to 4 cycles. During the first three cycles, patients were treated

with IUI; in the fourth cycle, IVF with maximal ovarian hyperstimulation was performed. The IVF-cycle had to be initiated within a year from the first cycle.
It seems that in IVF stimulation the maximal effect is reached with FSH dosages up to 225 IU per day (The Latin-American puregon IVF Study Group, 2001, Out et al., 2000, Out et al., 2001). Using these results, we concluded that an initial stimulation by 3 ampoules of 75 IU of FSH under a long (GnRH-agonist suppressed) protocol, probably gives a maximal IVF stimulation, the outcome in terms of number of follicles and oocytes of which could be used as the golden standard for the cohort size.

Clomiphene citrate challenge test (CCT): starting on the fifth day of the menstrual cycle (CD 1 = day of onset of menses) 100 mg of Clomiphene citrate (Serophene®; Serono, Geneva, Switzerland) was administered for 5 days. In this study on CD 2 or 3 (basal values) and on CD 10 (stimulated values) the serum FSH was determined. Analysis of the CCT (Kwee *et al.*, 2003) was performed by the parameter: $\sum$ bFSH + sFSH.

Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT): on CD 3, 300 IU recFSH (Gonal-F®, Serono, Geneva, Switzerland) were administered subcutaneously (s.c). In this study blood samples for the determination of FSH, E2 and Inhibin B were drawn: just before (basal values) and 24 hrs after (stimulated values) the administration of FSH. Analysis of the EFORT (Kwee *et al.*, 2003) included the following parameters: E2-increment and Inhibin B-increment 24 hrs after administration of FSH.
The bFSH level was determined as an integral part of all CCT's and EFORT's.
All tests (CCT and EFORT), during the first 3 test cycles were followed by regular IUI cycles. Each cycle was monitored with serial transvaginal sonography (TVS) to evaluate if clomiphene citrate caused multifollicular growth and what the effect of a single injection with 300 IU recFSH in the early follicular phase was. When the leading follicle reached a diameter of 18-20 mm (measured in two perpendicular directions), 10.000 IU of hCG (Profasi, Serono) was administered to induce final follicular maturation. The IUI was performed 42 hours later. When after 3 IUI-treatment cycles no pregnancy had occurred, an IVF-treatment cycle was the next step. The IVF-cycle had to be initiated within a year from the first CCT or EFORT.

IVF-treatment: The ovarian hyperstimulation protocol was performed according to a long GnRH-agonist protocol starting in the midluteal phase. On CD 3 of the first cycle the CCT or the EFORT was performed as described above. In the subsequent midluteal phase, seven days after ovulation, daily s.c. injections with triptoreline-acetate (Decapeptyl®, 0.1 mg/day; Ferring, Hoofddorp, the Netherlands) were started. Because of the administration of the GnRH-agonist, patients were advised to use a barrier type of contraception during this cycle. On CD 3 of the next cycle, ovarian hyperstimulation was started with daily s.c. injections of a fixed dose of 225 IU uFSH (Metrodin HP®, 75 IU/amp; Serono, Geneva, Switzerland), because this dosage probably gives a maximal effect in follicle stimulation. Standard procedures were followed including transvaginal sonography (TVS) (Aloka SSD-1700, 5.0 MHz probe) on CD 2 or 3 and on CD 9 or 10. Daily TVS was performed from the moment when the leading follicle reached a diameter of 16 mm. Ovarian hyperstimulation was continued until the largest follicle reached a diameter of at least 18 mm. The maximum duration of uFSH administration allowed was 16 days. If these criteria were met, Metrodin HP® and Decapeptyl® were discontinued and 10.000 IU of hCG (Profasi®, 10.000 IU/amp;

Serono, Geneva, Switzerland) were administered. On the day of hCG, TVS was performed to count the result of ovarian hyperstimulation (all follicles ≥ 10 mm) expressed as the total number of follicles.

Serum assay

Serum estradiol (E2) and FSH were determined by commercially available immunometric assays (Amerlite, Amersham, UK). For E2, the inter-assay CV was 11 % at 250 pmol/l and 8 % at 8000 pmol/l, the intra-assay coefficient of variation (CV) was 13 % at 350 pmol/l, 9 % at 1100 pmol/l and 9 % at 5000 pmol/l. The lower limit of detection for E2 was 90 pmol/l. In the EFORT and CCT we measured estradiol by a sensitive radioimmunoassay (Sorin, Biomedica, Saluggia, Italy). This measurement of estradiol was abbreviated as EE. For EE, the inter-assay CV was 11 % at 60 pmol/l, 8 % at 200 pmol/l, 11 % at 550 pmol/l and 8 % at 900 pmol/l. The intra-assay CV was 4 % at 110 pmol/l and 5 % at 1000 pmol/l. The lower limit of detection for EE was 18 pmol/l. For FSH, the inter-assay CV was 9 % at 3 IU/l and 5 % at 35 IU/l, the intra-assay CV was 9 % at 5 IU/l, 8 % at 15 IU/l and 6 % at 40 IU/l. The lower limit of detection for FSH was 0.5 IU/l. Inhibin B was determined immunometrically by a commercially available assay (Serotec Limited Oxford UK). For Inhibin B, the inter-assay CV was 17 % at 25 ng/L, 14 % at 55 ng/L and 9 % at 120 ng/L and the intra-assay CV was 8 % till 40 ng/l and 5 % at > 40 ng/l. The lower limit of detection for Inhibin B was 13 ng/l.

Half way through the study, the Amerlite assay used to assess FSH was suddenly withdrawn from the market and had to be replaced by another commercially available assay (Delfia, Wallac, Finland). The two assays have been compared and showed excellent linear correlation, although a shift in the values took place (Delfia FSH = 1.28 x Amerlite FSH + 0.01 (r=0.9964)). For the Delfia FSH, the inter-assay CV was 5 % at 3.5 IU/l and 3 % at 15 IU/l. All FSH determinations have been recalculated and are expressed according to the Delfia assay. Values below the detection limit of an assay were assigned a value equal to the detection limit of that assay.

Statistical analysis

Descriptive statistics and univariate tests were carried out using SPSS for Windows. Data on inter-cycle variability was analysed using random coefficient models which are a generalisation of ordinary regression models. Such models allow us to remove variation due to other factors, such as age. A brief description of random coefficient models and details of the models fitted are given in an addendum.

The measure of intercycle variation used was the within-patient standard deviation of each test variable over time. As a standard deviation can only be defined when at least two measurements have been done, patients with only one measurement point were excluded from the analysis.

Secondly, we examined the relationship between intercycle variation and ovarian reserve by calculating the correlation coefficient between variability and ovarian capacity. For the ovarian reserve, we used the result of ovarian hyperstimulation expressed as the total number of retrieved oocytes as golden standard. That means that only patients with 4 tests were included in this part of the analysis.

Results

The characteristics of the 2 groups are given as means ± SD in Table I. No significant differences were noted between the groups in baseline characteristics, cycle day 3 measurements or outcome parameters. In the CCT group, 69.8 % had a primary infertility and 30.2 % a secondary infertility. The cause of infertility was for 62.8 % an idiopathic factor, 30.2 % a male factor and 7.0 % a cervical factor. In the EFORT group, 66.7 % had a primary infertility and 33.3 % a secondary infertility. The cause of infertility was for 63.4 % an idiopathic factor, 31.7 % a male factor and 4.9 % a cervical factor.

Table I Characteristics of the groups (values are means ± SD). No significant differences.

	CCT-group N = 43	EFORT-group N = 42
Baseline characteristics		
Age (y)	33.0 ± 3.3	32.5 ± 3.7
Duration infertility (y)	3.8 ± 1.3	4.1 ± 1.5
Cycle 1 day 3		
FSH (IU/l)	7.0 ± 2.3	7.7 ± 2.2
E2 (pmol/l)	130.7 ± 59.2	121.2 ± 85.8
Inhibin B (ng/l)	108.0 ± 45.8	92.3 ± 54.2
Treatment results cycle 4	*N = 26*	*N = 33*
Duration of stimulation (d)	11.6 ± 1.6	11.9 ± 2.6
Number of ampoules of FSH	32.0 ± 4.7	32.8 ± 7.9
E2 level on the day of hCG (pmol/l)	15116.5 ± 24798.2	13848.2 ± 22688.9
Endpoints cycle 4		
Total number of follicles	14.3 ± 10.2	15.5 ± 10.5
Total number of oocytes	11.4 ± 7.6	12.4 ± 8.9

As shown in table II, in the CCT group, 6 patients received 1 test, 10 received 2 tests, 1 received 3 tests and 26 received 4 tests. Six patients conceived after the first test, 7 after the second test and 1 after the third. Three patients dropped out of the study for unknown reasons after 2 tests.

In the EFORT group, four patients became pregnant spontaneously after randomization and before starting with the IUI treatment. Two patients received 1 test, 1 received 2 tests, 2 received 3 tests and 33 received 4 tests. After the first test 1 patient conceived, another following the second and 2 after the third test. One patient became pregnant in the first month of IVF-treatment. One patient dropped out of the study for unknown reasons after one test.

To exclude bias in analysis of the patients groups due to pregnancy and drop out rates, we performed statistical analysis on the characteristics. No significant differences were noted between the two study groups during the study.

Table II Breakdown of number of patients that became pregnant or dropped out during the subsequent IUI and IVF treatments devided over the two test groups.

Total n = 85	Cycle 0	Cycle 1 – IUI			Cycle 2 - IUI			IUI Cycle 3 - IUI			Cycle 4 - IVF		
	Preg.[1]	N.	Preg.	Drop out	N.	Preg.	Drop out	N.	Preg.	Drop out	N.	Preg.	Drop out
CCT n = 43	0	43	6	0	37	7	3	27	1	0	26	4	1
EFORT n = 42	4	38	1	1	36	1	0	35	2	0	33	8	1

[1] Pregnant after randomization but before treatment

Pregnant = ongoing pregnancy defined as positive hartbeat by a gestational age of 12 weeks

The regression analysis revealed no systematic change in mean level of the clinical variables over time (table III). However, significant intercycle variation was seen in two of the four variables. The per-cycle variance of bFSH ranged from 1.8 to 4.4 (maximum to minimum ratio of 2.44, $p<0.0001$), while that of CCT ranged from 21.3 to 70.6 (3.31, $p<0.0001$). No significant change in per cycle variance was found for the E2-increment (1.25, $p>0.2$) and Inhibin B-increment (1.31, $p>0.2$).

Table III Per-cycle variance for each of the four variables.

Cycle	BFSH (IU/l)[1]	CCT (IU/l)[1]	EFORT	
			Inh B-incr. (ng/l)[2]	E2-incr. (pmol/l)[2]
1	2.1	21.3	2498	4668
2	4.4	70.6	3003	3746
3	3.0	40.6	2441	4632
4	1.8	40.6	2296	3767

[1]P<0.0001, [2]not significant

Table IV shows the correlation between ovarian reserve and intercycle variability. The intercycle variability of both CCT and basal FSH (Figure 1A and 1B) were strongly and negatively linked with ovarian reserve. This negative correlation indicates that women who have high intercycle variation in these factors tend to have a lower ovarian reserve. The intercycle variability of the E2-increment showed no apparent correlation with ovarian reserve while the intercycle variability of the Inhibin B-increment was positively correlated with ovarian reserve and was significant at the 6% level.

Table IV Correlation between ovarian capacity and inter-cycle variance.

	BFSH (IU/l)	CCT (IU/l)	EFORT	
			Inh B-incr. (ng/l)	E2-incr. (pmol/l)
R	-0.52	-0.54	0.337	0.194
p-value	0.008	0.000	0.060	0.286

Figure 1. (A) The distribution of the intercycle variability of bFSH (IU/l), expressed by the standard deviation according to the number of oocytes.

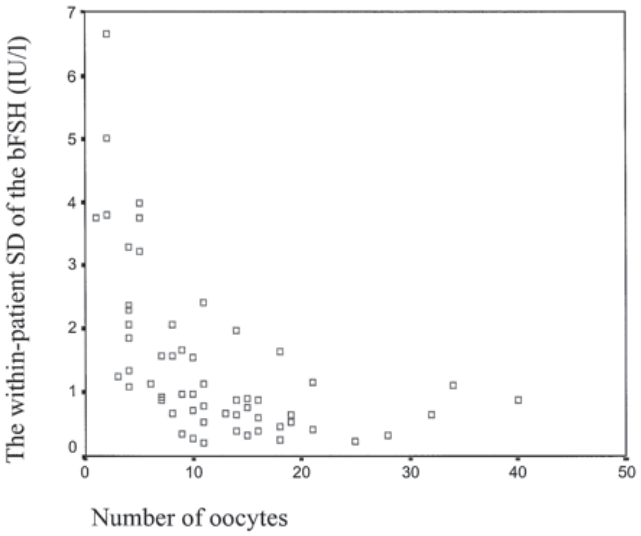
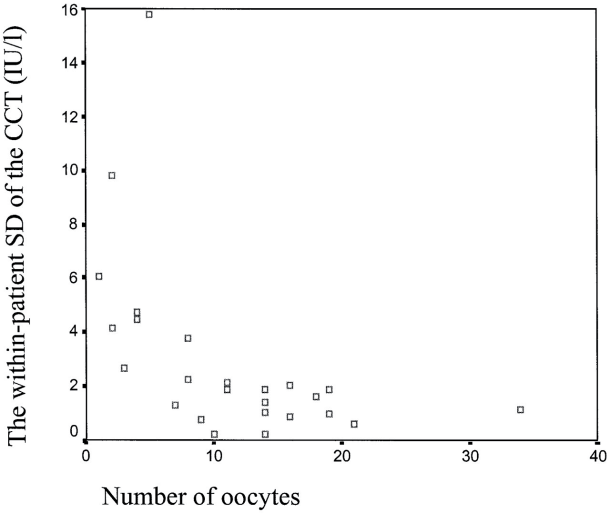


Figure 1. (B) The distribution of the intercycle variability of CCT (IU/l), expressed by the standard deviation according to the number of oocytes



Discussion

Recently, we showed that the inhibin-B increment and E2-increment in the EFORT were the best predictors of the total number of follicles obtained after maximal ovarian hyperstimulation in an IVF treatment, i.e. cohort size (Kwee *et al.*, 2003). Age, basal values of FSH, E2 and Inhibin B and the outcome of the CCT in this respect, each, and in combination, showed a much lower performance. In order to allow correct interpretation of these tests, we now estimated their intercycle variability.

This study shows that the intercycle variability of the Inhibin-B and estradiol increment in the EFORT is not significantly different in consecutive cycles. This, in contrast to bFSH and CCT which vary significantly from cycle to cycle. This outcome indicates that the EFORT potentially is a highly reproducible predictor for ovarian reserve.

The issue of intercycle variability of ovarian reserve tests as a marker for ovarian reserve has been studied in the past (Brown *et al.*, 1995, Hannoun *et al.*, 1998, Scott *et al.*, 1990). In 1990, Scott *et al*. evaluated and documented the intercycle variation of two cycles of basal FSH (Scott *et al.*, 1990). They found that two basal FSH values that are in agreement might be used to counsel patients regarding their performance during hyperstimulation. Women with a normal bFSH had a small range in the intercycle variation, in contrast to women with an elevated bFSH showing a much greater variation. If the patient had wide fluctuations in her basal FSH values, she was more likely to respond poorly to hyperstimulation. They suggested therefore serial screening of bFSH because they found the diagnostic and predictive value of a normal value of a single determination of bFSH limited. Brown *et al.* (Brown *et al.*, 1995) evaluated the intercycle variation of bFSH in a group of normally cycling women, unselected for fertility. They concluded that a single day 3 FSH level, <20 IU/L is highly predictive of all subsequent values within a year in women under 40. Women over 40 years with a 'normal' day 3 level had a 50% chance of having an elevated day 3 FSH level in the subsequent cycle. In this group the determination of more than one determination of a day 3 FSH level was likely to have prognostic significance. Hannoun *et al.* (Hannoun *et al.*, 1998) documented the variation of the results of the CCT performed in the same patient from cycle to cycle. They showed a high degree of intercycle variability. But they did not test the influence of this variability on the ovarian reserve.

Our study confirms that the intercycle variation of bFSH and CCT was strongly negatively correlated with the outcome of IVF. A small intercycle variability of bFSH and CCT was associated with a 'normal' ovarian reserve. Patients with large fluctuating bFSH and CCT results in consecutive cycles had fewer follicles after controlled hyperstimulation. This confirms that high intercycle variability of bFSH and CCT could act as a marker for low response as result from limited ovarian reserve. However, from our study it appeared that intercycle variability of bFSH and CCT outcome has no added value to a single measurement for prediction of ovarian reserve. This is probably because elevated basal FSH and abnormal outcome of CCT are coupled to the phenomenon of high intercycle variability. Women with one single 'elevated' bFSH show a high intercycle variability and have a high chance for a 'poor' response after ovarian hyperstimulation and women with one single 'normal' bFSH show a small intercycle variability and have a high chance for an 'adequate' response after ovarian hyperstimulation (Scott *et al.*, 1990). Comparing the prediction of intercycle variability of bFSH and CCT with one single screening of bFSH and CCT (Kwee *et al.*, 2003) there was no added value of the intercycle variability for the prediction of ovarian reserve.

Therefore it seems that there is no need for evaluation of intercycle variability of bFSH and CCT as a marker for ovarian reserve.
From a pathophysiological point of view, large intercycle variability in bFSH and CCT, particularly in those patients with limited ovarian reserve remains an intruiging phenomenon. Early follicular phase fluctuations in FSH are a reflection of the balance between ovarian steroid and peptide inhibition and the hypothalamic pituitary drive during the period of follicular recruitment at the moment just before the selection of the dominant follicle. It signifies the amount of inhibin and/or E2 produced by the cohort of follicles, responsible for the negative feedback on the FSH secretion. The bFSH increases when ovarian reserve diminishes (Lenton *et al.*, 1988), supposedly because the small antral follcles produce less inhibin B and possibly E2.
De Koning *et al.* (de Koning *et al.*, 2000) found that elevated basal FSH results from a more sensitive pituitary to GnRH, leading to higher FSH and LH pulses. So, this means that FSH in the early follicular phase as well as after stimulation with clomiphene is influenced in many ways and is potentially susceptible to large variations particularly with a small follicle cohort. Therefore basal FSH and CCT can only act as a very indirect measure for the actual cohort size.
Contrastingly, why is the intercycle variability of the EFORT very small? EFORT is a very accurate predictor of ovarian response. Therefore, a likely explanation would be that the monthly available ovarian follicle cohort is surprisingly constant. Even in women with a small cohort. In about 70 days of follicular development, the primary follicles reach the early antral stage. They become sensitive to FSH (Gougeon *et al.*, 1996). At the end of the first 70 days (Scheele *et al.*, 1996, Schoemaker *et al.*, 1993, Van der Meer *et al.*, 1994), in the early follicular phase, a cohort of antral follicles, probably approximately 20 in number, each with a different sensitivity to FSH, is present in the ovaries, ready to continue their development under the stimulation of FSH. This is the stage in which we perform the EFORT. The EFORT is a direct test, which demonstrates the ability of the ovaries to initiate aromatase activity in response to a fixed dose of exogenously administered FSH (300 IU FSH). The aromatase activity results in increased follicular concentrations of estradiol since the aromatase substrate, androstenedione is already available in abundance. Inhibin B is produced by the granulosa cells of the developing cohort of follicles and therefore directly reflects the ovarian capacity (Groome *et al.*, 1996, Welt *et al.*, 1997). Apparently the month to month variability of the cohort of follicles measuring 2-5 mm, present very early in the follicular phase of the cycle (Chang *et al.*, 1998) is very small.
Indeed, Scheffer *et al.* (Scheffer *et al.*, 1999) using sonography, found a limited intercycle variability of antral follicle count (AFC) in regularly cycling women. Sofar, no other studies have been published that describe intercycle variability of AFC.
Similarly, no studies have currently been published on intercycle variablity of other available tests for ovarian reserve such as GnRH-agonist stimulation test (GAST) (Padilla *et al*., 1990, Ravhon *et al.*, 2000) and the measurement of Anti- Müllerian Hormone (AMH) (Vet *et al.*, 2002).
Ideally, a test for ovarian reserve should be short, simple, accurate and precise. We conclude that the EFORT in comparison to bFSH and CCT has superior stable characteristics in terms of intercycle variability. This high reproducibility and its optimal performance to predict ovarian response (Kwee *et al.*, 2003) makes the EFORT a useful endocrine test to predict ovarian reserve.

Addendum

Random coefficient models are a generalisation of ordinary regression models. Consider the following regression model

$$y = \alpha + \beta x + \varepsilon$$

where y is the dependent variable, x is an explanatory variable, α and β are unknown coefficients and ε is the random error. Usually, the coefficients α and β are considered as being fixed parameters, taking the same value for all individuals. However, in some situations, one or other of the coefficients (typically the coefficient associated with the explanatory variable) may vary from individual to individual. We may therefore choose to view β as being a random variable with a given distribution (usually normal) with a fixed mean $\beta 0$ and variance $\sigma\beta 2$ i.e. $\beta \sim N(\beta 0, \sigma\beta 2)$.

Alternatively, we can write $\beta = \beta 0 + \beta 1$ where $\beta 0$ is fixed and $\beta 1 \sim N(0, \sigma\beta 2)$.

Entering this expression into the regression model gives

$$y = \alpha + (\beta 0 + \beta 1)x + \varepsilon = \alpha + \beta 0x + (\varepsilon + \beta 1x)$$

Since $\beta 1$ is random, the term in brackets is random and represents the random variation in the data. Usually, $\beta 1$ is assumed to be uncorrelated to ε.

The differences between the models are twofold. First, the observed variation is split into two components. Secondly, the amount of variation is to some degree dependent on the covariate x.

A hypothesis test of no x-effect on the variation (i.e. H0: $\sigma\beta 2=0$) can be carried out the values of the log-likelihood functions of two models, one containing the extra term, the other without it.

In our analyses, we defined three binary variables, T2, T3 and T4. T2 took the value one if an observation corresponded to the second cycle and zero if it did not. The other variables were defined analogously. Our model was therefore

$$y = \alpha + \Sigma\beta 0iTi + \gamma(age) + (\varepsilon + \Sigma\beta 1iTi)$$

where the summation is over the T variables. Note that we are not really interested in the coefficients $\{\beta 0i\}$ but we have included them in the analyses.

Acknowledgements
The authors acknowledge the help of Dr Corry Popp-Snijders and her staff, particularly for the endocrine laboratory work and the staff of the IVF centre for assistance during the execution of the protocol. This study was financially supported by Serono, Geneva, Switzerland.

References

Brown JR, Liu HC, Sewitch KF, Rosenwaks Z, Berkeley AS. Variability of day 3 follicle-stimulating hormone levels in eumenorrheic women. J Reprod Med 1995;40: 620-4.

Chang MY, Chiang CH, Hsieh TT, Soong YK, Hsu KH. Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil Steril 1998;69:505-10.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A, Frydman R. Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Human Reprod 1994;9:1607-11.

Gougeon A. Regulation of ovarian follicular development in primates: facts and hypotheses. Endocr Rev 1996;17:121-55.

Groome NP, Illingworth PJ, O'Brien M, Pai R, Rodeger FE, Mather JP, Mcneilly AS. Measurement of dimeric inhibin B throughout the human menstrual cycle. J Clin Endocrinol Metab 1996;81:1401-5.

Hannoun A, Abu Musa A, Awwad J, Kaspar H, Khalil A. Clomiphene citrate challenge test: cycle to cycle variability of cycle day 10 follicle stimulating hormone level. Clin Exp Obstet Gyn 1998;25:155-6.

Koning de CH, Popp-Snijders C, Schoemaker J, Lambalk CB. Elevated FSH concentrations in imminent ovarian failure are associated with higher FSH and LH pulse amplitude and response to GnRH. Hum Reprod 2000;15:1452-6.

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB, Schoemaker J. Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study. Hum Reprod 2003;18:1422-7.

The Latin-American Puregon IVF Study Group. A double-blind clinical trial comparing a fixed daily dose of 150 and 250 IU of recombinant follicle-stimulating hormone in women undergoing in vitro fertilization. Fertil Steril 2001;76: 950-6.

Lenton EA, Sexton L, Lee S, Cooke ID. Progressive changes in LH and FSH and LH: FSH ratio in women throughout reproductive life. Maturitas 1998;10: 35-43.

Navot D, Rosenwaks Z, Margalioth EJ. Prognostic assessment of female fecundity. Lancet 1987;2: 645-7.

Out HJ, Braat DM, Lintsen BME, Gurgan T, Bukulmez O, Gokmen O, Keles G, Caballero P, Gonzalez JM, Fabregues F, Balasch J, Roulier R. Increasing the daily dose of recombinant follicle stimulating hormone (Puregon®) does not compensate for the age-related decline in retrievable oocytes after ovarian stimulation. Hum Reprod 2000;15: 29-35.

Out HJ, David I, Ron-El R, Friedler S, Shalev E, Geslevich J, Dor J, Shulman A, Ben Rafael Z, Fisch B, Dirnfeld M. A randomized, double-blind clinical trial using fixed daily dose of 100 or 200 IU of recombinant FSH in ICSI cycles. Hum Reprod 2001;16:1104-9.

Padilla SL, Bayati J, Garcia JE. Prognostic value of the early serum estradiol response to leuprolide acetate in in vitro fertilization. Fertil Steril 1990;53:288-94.

Ravhon A, Lavery S, Michael S, Donaldson M, Margara R, Trew G Winston R. Dynamic assays of inhibin B and oestradiol following buserelin acetate administration as predictors of ovarian response in IVF. Hum Reprod 2000;15:2297–301.

Scheele F, Schoemaker J. The role of follicle-stimulating hormone in the selection of follicles in human ovaries: a survey of the literature and a proposed model. Gynecol Endocrin 1996;10:55-66.

Scheffer GJ, Broekmans FJ, Dorland M, Habbema JD, Looman CW, te Velde, E.R. Antral follicle counts by transvaginal ultrasonography are related to age in women with proven natural fertility. Fertil Steril 1999;72:845–51.

Schoemaker J, van Weissenbruch MM, Scheele F, van de Meer M. The FSH threshold concept in clinical ovulation induction. Bail Clin Obstet Gynecol 1993;7:297-308.

Scott RT, Toner JP, Muasher SJ, Oehninger S, Robinson S, Rosenwaks Z. Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. Fertil Steril 1989;51: 651-4.

Scott RT, Hofmann GE, Oehninger S, Muasher SJ.Intercycle variability of day 3 follicle-stimulating hormone levels and its effect on stimulation quality in in vitro fertilization. Fertil. Steril 1990;54:297-302.

Sharara FI, Scott RT, Seifer DB. The detection of diminshed ovarian reserve in infertile women. Am J Obstet Gynecol 1998;179:804-12.

Van der Meer M, Hompes PGA, Scheele F, Schoute E, Veersema S, Schoemaker J. Follicle stimulating hormone (FSH) dynamics of low dose step-up ovulation induction with FSH in patients with polycystic ovary syndrome. Hum Reprod 1994;9:1612-7.

Vet A, Laven JSE, de Jong FH, Themmen APN, Fauser BCJM. Antimüllerian hormone serum levels: a putative marker for ovarian aging. Fertil Steril 2002;77:357-62.

Welt CK, Martin KA, Taylor AE, Lambert-Messerlian GM, Crowley WF, Smith JA, Schoenfeld DA, Hall, JE. Frequency modulation of follicle- stimulating Hormone (FSH) during the luteal-follicular transition: Evidence for FSH control of inhibin B in normal women. J Endocrinol Metab 1997;82:2645-52.

Letter to the editor: Variability of ovarian reserve tests

Koray Elter, Alper Sismanoglu and Fatih Durmusoglu
Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, Marmara University School of Medicine, Istanbul, Turkey
Human Reproduction 2004;19:2170

Sir,
The article by Kwee *et al.* (2004) on the intercycle variability of ovarian reserve tests contains important methodological points that require further explanation and clarification before valid conclusions can be drawn.

i. The authors mention that cycle day 2 or 3 serum FSH values were determined as basal values during clomiphene citrate challenge test (CCCT). It has been reported that there is considerable variation in serum FSH levels between days 2 and 3 (Brown *et al.*, 1995). The significant variation in that study (Kwee *et al.*, 2004) may partly be due to this intra-cycle variability. Therefore, we believe that the results on the intercycle variability of CCCT in that study should be cautiously interpreted.

ii. In the relevant study, three ovarian reserve tests have been performed one to four times in subsequent cycles. Although it has not been mentioned whether subjects were on any recent treatment, i.e. ovulation induction, prior to enrollment, we assume that subjects did not have any ovulation induction before the first cycle. However, ovarian reserve tests in cycles 2, 3 and 4 were performed after ovulation induction. To our knowledge, the effect of clomiphene citrate on the ovarian reserve tests in the following cycle has not been discovered. It has been reported that significant plasma concentrations of clomiphene citrate could be detected up to one month after treatment with a single dose of 50 mg (Mikkelson *et al.*, 1986). Table III shows that the major variation was observed between cycles 1 and 2, i.e. between the cycle following a spontaneous cycle and that following an ovulation induction cycle. Any possible effect of clomiphene citrate may be responsible for that variability. It may be more appropriate to analyse the intercycle variability between similar cycles, i.e. either between cycles with prior ovulation induction or between cycles without any prior treatment, and it may also be appropriate to exclude cycle 1 from the analysis in the relevant study.

iii. To exclude the bias of pregnant subjects during the study, the authors mention that the CCCT and exogenous FSH ovarian reserve test (EFORT) groups were comparable. However, the basis of the study is the variability between cycles. It is obvious that subjects in each cycle are different due to pregnancies. It would also be inconclusive to compare characteristics of pregnant subjects with those of others to exclude any bias, due to the type II error. Due to the small number of pregnant subjects, it would not mean that pregnant subjects were comparable to others if the statistical analysis could not reveal any significance. The bias of pregnant subjects may have altered the results in the study.

iv. The authors mention that there is significant intercycle variability in basal FSH and CCCT values based on the results, which were shown in Table III. Variance is: Σ(value – mean)2/(n – 1) and SD is the square root of the variance. To our understanding, variances in cycles 1 to 4 (per cycle variation) have been compared and this has been reported as the intercycle variability. However, since the populations in each cycle are different due to pregnancies, this comparison is not appropriate.

It may be more appropriate to calculate the variance per subject, i.e. variance could be calculated for the values of the same subject in subsequent cycles. That is how inter-assay variabilities are calculated: a constant serum sample is tested multiple times at different times and the variance of these values indicates the inter-assay variability. The intra-assay variation describes the variation between multiple assay wells on the same plate from the same sample. It is our understanding that variances of populations were compared in the relevant study, instead of variances of ovarian reserve tests in the same subject.

Variance of X1, Y1 and Z1 values are relevant for the intercycle variability. Variance of X1, X2 and X3 values indicate the population variances. Similar to the constant use of the same serum sample for analysis in the example of assay variability, subjects should be constant in each cycle. Otherwise, comparison of variances may mislead data, and any significant variability may be due to the inequality of the population.

	Cycle 1	Cycle 2	Cycle 3
Subject 1	X1	Y1	Z1
Subject 2	X2	Y2	Z2
Subject 3	X3	Y3	Z3

References

Brown JR, Liu HC, Sewitch KF, Rosenwaks Z and Berkeley AS (1995) Variability of day 3 follicle-stimulating hormone levels in eumenorrheic women. J Reprod Med 40, 620–624.

Kwee J, Schats R, McDonnell J, Lambalk CB and Schoemaker J (2004) Intercycle variability of ovarian reserve tests: results of a prospective randomized study. Hum Reprod 19, 590–595.

Mikkelson TJ, Kroboth PD, Cameron WJ, Dittert LW, Chungi V and Manberg PJ (1986) Single-dose pharmacokinetics of clomiphene citrate in normal volunteers. Fertil Steril 46, 392–396.

Reply: Variability of ovarian reserve tests

J. Kwee, R. Schats, J. McDonnell, C.B. Lambalk and J. Schoemaker
Division of Reproductive Endocrinology and Fertility and the IVF Centre, Department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands
Human Reproduction 2004;19:2171

Sir,
We thank Dr Elter *et al.* for their comments concerning our paper on intercyclical variability of ovarian reserve tests.
The authors raise the issue of possible differences between day 2 and day 3 FSH values which in our study may have contributed to the observed variation of FSH between cycles. In the first place, only in ~10% of the participants was FSH not measured on all third days of the cycle. The cited small subanalysis in 20 patients by Brown *et al.* (1995). indicated a possible within-cycle coefficient of variation for FSH of 14.8% It should be realized that such a variation also included the assay varation (4.8% intra-assay variation and 6.2% inter-assay variation) and that value indicates that the within (intra)-cycle variation of FSH measurement is probably only limited. And indeed Hansen et al. (1996), in the study that we cited, measured FSH on cycle days 2–5 in order to investigate the intra- and intercycle variability in a healthy population of 44 women with regular menstrual intervals in a total of 66 cycles on cycle days 2, 3, 4 and 5, and FSH concentrations were not different between the various cycle days.
The second point raised was about the possible carry-over effect of clomiphene from one cycle to the other. Indeed, it has been reported that significant plasma concentrations of clomiphene citrate could be detected up to 1 month after treatment with a single dose of 50 mg (Mikkelson *et al.*, 1986). But this is predominantly the so-called isomeric Zu variant of clomiphene. Glasier *et al.* (1989) investigated the effects on follicular development of clomiphene citrate and its two isomers En clomiphene and Zu clomiphene. It was concluded that the En isomer, which has largely the antiestrogenic properties, is the isomer active in inducing follicular development. The biologically active En clomiphene is elimated much more quickly than the biologically inactive Zu clomiphene. Moreover, Opsahl *et al.* (1996) showed that patterns of gonadotrophin response, follicular development, and endometrial growth and maturation remain consistent across consecutive cycles of clomiphene citrate treatment. This is why we believe that there is no carry-over effect of clomiphene citrate. Taking this altogether, we assumed that a biological carry-over effect of clomiphene in our study could be negligible.
The third point raised was whether the bias of pregnant subjects may have biased the results in the study. Indeed we do have a potential bias here in that women who became pregnant during the three test cycles did not reach the IVF cycle in which the ovarian reserve was evaluated. We admit the possibility of this bias but we could not think of a way to avoid it in an ethically acceptable manner. It turned out that the number of pregnant subjects was (relatively) small. Therefore this bias, if present, has only contributed to a limited extent.
Finally, it was suggested that we should have calculated the variance per subject, i.e. variance within-subject over cycles. This is in fact exactly what we have done: SD = square root of variance measured within each female patient.

References

Brown JR, Liu HC, Se Witch KF, Rosenwaks Z and Berkeley AS (1995) Variability of day 3 follicle-stimulating hormone levels in eumenorrheic women. J Reprod Med 40, 620–624.

Glasier AF, Irvine DS, Wickings EJ, Hillier SG and Baird DT (1989) A comparison of the effects on follicular development between clomiphene citrate, its two separate isomers and spontaneous cycles. Hum Reprod 4, 252–256.

Hansen LM, Batzer FR, Gutmann JN, Corson SL, Kelly MP and Gocial B (1996) Evaluating ovarian reserve: follicle stimulating hormone and oestradiol variability during cycle days 2-5. Hum Reprod 11, 486–489.

Mikkelson TJ, Kroboth PD, Cameron WJ, Dittert LW, Chungi V and Manberg PJ (1986) Single-dose pharmacokinetics of clomiphene citrate in normal volunteers. Fertil Steril 46, 392–396.

Opsahl MS, Robins ED, O'Connor DM, Scott RT and Fritz MA (1996) Characteristics of gonadotropin response, follicular development, and endometrial growth and maturation across consecutive cycles of clomiphene citrate treatment. Fertil Steril 66, 533–539.

Chapter 5

Ovarian volume and antral follicle count for the prediction of low and hyper responders with in vitro fertilization

J. Kwee, M.W. Elting, R. Schats, J. McDonnell and C.B. Lambalk

Division of Reproductive Endocrinology and Fertility and the IVF Centre, department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.

Submitted

Abstract

Background: The current study was designed to compare antral follicle count (AFC) and basal ovarian volume (BOV), the exogenous FSH ovarian reserve test (EFORT) and the clomiphene citrate challenge test (CCCT), with respect to their ability to predict poor and hyper responders.

Methods: One hundred and ten regularly menstruating patients, aged 18-39 years, participated in this prospective study, randomized, by a computer designed 4-blocks system study into two groups. Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. All patients underwent a transvaginal sonography to measure the basal ovarian volume and count of basal antral follicle. In all patients, the test was followed by a standard IVF treatment. The result of ovarian hyperstimulation during IVF treatment, expressed by the total number of follicles, was used as gold standard.

Results: The best prediction of ovarian reserve (Y) was seen in a multiple regression prediction model that included, AFC, Inhibin B-increment in the EFORT and BOV simultaneously (Y = -3.161 + 0.805 x AFC (0.258-1.352) + 0.034 x Inh. B-incr.(0.007-0.601) + 0.511 BOV (0.480-0.974)($r=0.848$, $p<0.001$). Univariate logistic regression showed that the best predictors for poor response were the CCCT (ROC-AUC = 0.87), the bFSH (ROC-AUC = 0.83) and the AFC (ROC-AUC = 0.83). Multiple logistic regression analysis did not produce a better model in terms of improving the prediction of poor response. For hyper response, univariate logistic regression showed that the best predictors were AFC (ROC-AUC = 0.92) and the inhibin B-increment in the EFORT (ROC-AUC = 0.92), but AFC had better test characteristics, namely a sensitivity of 82 % and a specificity 89 %. Multiple logistic regression analysis did not produce a better model in terms of predicting hyper response.

Conclusions: In conclusion AFC performs well as a test for ovarian response being superior or at least similar to complex expensive and time consuming endocrine tests. It is therefore likely to be the test for general practise.

Key Words: CCCT/EFORT/antral follicle count/ basal volume of the ovaries/ovarian reserve

Introduction

Real time two-dimensional (2D) pelvic ultrasonography is a relatively accurate and reliable method of determining ovarian volume and morphology (Campbell *et al.*, 1982). Interobserver and intraobserver measurements have been shown to be very low when using transvaginal sonography (Higgins *et al.*, 1990, Scheffer *et al.*, 2003).
The mean ovarian volume increases from 0,7 ml at 10 years to 5,8 ml at 17 years of age (Ivarsson *et al.*, 1983). It has been suggested that there are no major changes in ovarian volume during reproductive years until the premenopausal period. In women > 40 years old, there is a dramatic drop in ovarian volume, which is not related to parity (Andolf *et al.*, 1987, Higgins *et al.*, 1990, Ivarsson *et al.*, 1983). Thereafter there is a further sharp decline in size in postmenopausal women which seems mostly related to the time when menstruation ceases, rather than merely to age, because when oestrogen treatments were given, there appeared to be no decrease in ovarian volume with age (Andolf *et al.*, 1987).
Several studies (Syrop *et al.*, 1995, Tomás *et al.*, 1997, Bancsi *et al.*, 2004) demonstrate that ovarian volume, as determined by transvaginal ultrasonography, is a predictor of ovarian reserve and clinical pregnancy rate. Lass *et al.* (1997) confirmed that decrease in ovarian volume is an early sign of depletion of the follicles and its measurement is likely to be clinically useful.
A cohort of follicles measuring 2-5 mm is present very early in the follicular phase of the cycle (Chang *et al.*, 1998). These follicles are in an early antral phase, and are easily detected by transvaginal ultrasound, as they contain a small amount of antral fluid. The number of small follicles at the beginning of the cycle may well represent the actual functional ovarian reserve. So the number of small antral follicles are clearly related to age and could well reflect the size of the remaining primordial pool in women with proven natural fertility (Scheffer *et al.*, 1999, Kline *et al.*, 2005).
Previously (Kwee *et al.*, 2003), we published the comparison of endocrine tests for the prediction of the total number of follicles obtained after stimulation. With linear regression analysis, Inhibin B-increment and E2-increment in the EFORT gave the best predictive values. We tried to find one single, simple test, which could identify poor, normal and hyper responders (Kwee *et al.*, 2006) and concluded that by logistic regression analysis, the $\sum$ bFSH + sFSH in the CCCT was the best endocrine test to predict poor responders, unfortunately not for the prediction of hyper responders. The aim of the current study was to compare the antral follicle count (AFC) and the basal ovarian volume (BOV), with the exogenous FSH ovarian reserve test (EFORT) and the clomiphene citrate challenge test (CCCT), with respect to their ability to predict poor and hyper responders.

Materials and Methods

Study Population

One hundred and ten patients, aged 18-39 years, who were eligible for treatment by Intra Uterine Insemination (IUI) between june 1997 to december 1999 participated in the study. This study is part of a prospective randomized study of regular menstruating patients to the determination of ovarian reserve (Kwee *et al.*, 2003). Their infertility was either idiopathic for > 3 years and/or due to a male factor and/or cervical hostility. Cervical hostility was

diagnosed by means of a well timed negative postcoital test, that is, no progressive motile spermatozoa seen at a magnification of 400 x in good cervical mucus despite normal semen parameters.
Patients had to have regular menstrual cycles, two ovaries and showed two patent tubes with hysterosalpingography or at least one patent Fallopian tube with no further pathology with diagnostic laparoscopy. They were naive for IVF treatment. Excluded were patients with an oligo- or amenorrhoea (9 or fewer cycles a year) or a severe male factor, defined as 1. less than 1 million motile spermatozoa after Percoll centrifugation (gradient 40/90) and/or 2. > 20 % antibodies present on the spermatozoa after processing with Percoll centrifugation (gradient 40/90) and/or 3. > 50 % of the spermatozoa without an acrosome. Other exclusion criteria were untreated or insufficiently corrected endocrinopathies, clinically relevant systemic diseases or a body mass index > 28 kg/m².
The protocol was approved by the Institutional review Board and the Committee on ethics of research involving human subjects of the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands. All the couples participating in the study signed informed consent.

Treatment protocol
Patients were randomized by a computer designed 4-blocks system into two groups (Kwee *et al*, 2003). Fifty six patients underwent an transvaginal sonography to measure the basal ovarian volume and count of basal antral follicle and a Clomiphene citrate challenge test (CCCT), and 54 patients underwent an transvaginal sonography to measure the basal ovarian volume and count of basal antral follicle and an Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT). In all patients, the test was followed by an IVF treatment under a long protocol. The bFSH level, bE2 level and bInhibin B level were determined as an integral part of all CCCT's and EFORT's, as described previously (Kwee *et al.*, 2003). Van der Meer *et al.* (1998) showed that in eumenorrheic patients, the median (range) FSH threshold level for monofollicular growth was 5.3 (4.3-8.2) IU/l and the median (range) threshold dose was 75 IU (0.5-1.75) FSH/day. It was concluded that by an increment of ½ ampoule of FSH (37.5 IU) above the threshold dose for monofollicular growth, the maximum response is already obtained. It seems that in IVF stimulation maximal effect is reached with FSH dosages up to 225 IE (The Latin-American puregon IVF Study Group, 2001, Out et al., 2000, Out et al., 2001). Combining these facts, it can be concluded that an initial stimulation by 3 ampoules of 75 IU of FSH under a long (GnRH agonist suppressed) protocol, probably gives a maximal IVF stimulation, the outcome of which could be used as the gold standard for the cohort size.

Transvaginal Sonography Measurements
All ultrasound examinations were performed by one of the authors (J.K, R.S) using an Aloka SSD-1700 ultrasound apparatus (5.0 MHz probe).
The volume of each ovary was calculated by measuring in three perpendicular directions and applying the formula for an ellipsoid: (D1 x D2 x D3 x π/ 6). The volumes of both ovaries were added for the total basal ovarian volume (BOV).
To determine the diameter of the follicle, the mean of measurements in two perpendicular directions was taken. The numbers of follicles in both ovaries were added for the total antral follicle count (AFC). The follicles visualized and counted by TVS in the early follicular phase are 2-10 mm in size.

Clomiphene citrate challenge test (CCCT): starting on the fifth day of the menstrual cycle (CD 1 = day of onset of menses) 100 mg of Clomiphene citrate (Serophene®; Serono, Geneva, Switzerland) was administered for 5 days. In this study on CD 2 or 3 (basal values) and on CD 10 (stimulated values) the serum FSH was determined. Analysis of the CCCT (Kwee *et al.*, 2003) was performed by the parameter: $\sum$ bFSH + sFSH.

Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT): on CD 3, 300 IU recFSH (Gonal-F®, Serono, Geneva, Switzerland) were administered subcutaneously (s.c). In this study blood samples for the determination of FSH, E2 and Inhibin B were drawn: just before (basal values) and 24 hrs after (stimulated values) the administration of FSH. Analysis of the EFORT (Kwee *et al.*, 2003) included the following parameters: E2-increment and Inhibin B-increment 24 hrs after administration of FSH.

IVF-treatment: The ovarian hyperstimulation protocol was performed according to a long GnRH-agonist protocol starting in the midluteal phase. On CD 3 of the first cycle the ovarian volume and antral follicle count was measured by transvaginal sonography (TVS) examinations as described above. Also on CD 3 the CCCT or the EFORT was performed as described above. In the subsequent midluteal phase, seven days after ovulation, daily s.c. injections with triptoreline-acetate (Decapeptyl®, 0.1 mg/day; Ferring, Hoofddorp, the Netherlands) were started. Because of the administration of the GnRH-agonist, patients were advised to use a barrier type of contraception during this cycle. On CD 3 of the next cycle, ovarian hyperstimulation was started with daily s.c. injections of a fixed dose of 225 IU uFSH (Metrodin HP®, 75 IU/amp; Serono, Geneva, Switzerland), because this dosage probably gives a maximal effect in follicle stimulation. Standard procedures were followed including transvaginal sonography (TVS) (Aloka SSD-1700, 5.0 MHz probe) on CD 2 or 3 and on CD 9 or 10. Daily TVS was performed from the moment when the leading follicle reached a diameter of 16 mm. Ovarian hyperstimulation was continued until the largest follicle reached a diameter of at least 18 mm. The maximum duration of uFSH administration allowed was 16 days. If these criteria were met, Metrodin HP® and Decapeptyl® were discontinued and 10.000 IU of hCG (Profasi®, 10.000 IU/amp; Serono, Geneva, Switzerland) were administered. On the day of hCG, TVS was performed to count the result of ovarian hyperstimulation (all follicles $\geq$ 10 mm) expressed as the total number of follicles. TVS guided follicular aspiration (FA) was performed 36 hours after hCG administration. On the day of hCG administration E2 was determined. Follicular aspiration was done under systemic analgesia (7.5 mg diazepam orally and 50-100 mg pethidine hydrochloride intramuscularly), and all follicles present were aspirated.

Serum assay

Serum estradiol (E2) and FSH were determined by commercially available immunometric assays (Amerlite, Amersham, UK). For E2, the inter-assay CV was 11 % at 250 pmol/l and 8 % at 8000 pmol/l, the intra-assay coefficient of variation (CV) was 13 % at 350 pmol/l, 9 % at 1100 pmol/l and 9 % at 5000 pmol/l. The lower limit of detection for E2 was 90 pmol/l. In the EFORT and CCT we measured estradiol by a sensitive radioimmunoassay (Sorin, Biomedica, Saluggia, Italy). This measurement of estradiol was abbreviated as EE. For EE, the inter-assay CV was 11 % at 60 pmol/l, 8 % at 200 pmol/l, 11 % at 550 pmol/l and 8 % at 900 pmol/l. The intra-assay CV was 4 % at 110 pmol/l and 5 % at 1000 pmol/l. The lower

limit of detection for EE was 18 pmol/l. For FSH, the inter-assay CV was 9 % at 3 IU/l and 5 % at 35 IU/l, the intra-assay CV was 9 % at 5 IU/l, 8 % at 15 IU/l and 6 % at 40 IU/l. The lower limit of detection for FSH was 0.5 IU/l. Inhibin B was determined immunometrically by a commercially available assay (Serotec Limited Oxford UK). For Inhibin B, the inter-assay CV was 17 % at 25 ng/L, 14 % at 55 ng/L and 9 % at 120 ng/L and the intra-assay CV was 8 % till 40 ng/l and 5 % at > 40 ng/l. The lower limit of detection for Inhibin B was 13 ng/l.
Half way through the study (after 62 patients), the Amerlite assay used to assess FSH was suddenly withdrawn from the market and had to be replaced by another commercially available assay (Delfia, Wallac, Finland). The two assays have been compared and showed excellent linear correlation, although a shift in the values took place (Delfia FSH = 1.28 x Amerlite FSH + 0.01 (r=0.9964)). For the Delfia FSH, the inter-assay CV was 5 % at 3.5 IU/l and 3 % at 15 IU/l. All FSH determinations have been recalculated and are expressed according to the Delfia assay. Values below the detection limit of an assay were assigned a value equal to the detection limit of that assay.

Statistical analysis
The outcome measure of the first part of this study was the result of ovarian hyperstimulation expressed as the number of follicles. In our former study (Kwee *et al.*, 2003), we estimated the value of the independent variables by univariate linear regression, age, bFSH, CCCT-results, E2-increment in EFORT, inhibin B-increment in EFORT. In this study, we estimated by univariate linear regression, the value of the independent variables: total basal ovarian volume and the total basal antral follicle count in predicting the ovarian response. Stepwise regression analysis was used to find a prediction model for the ovarian response. The R square of the correlation of these variable(s) with the total number of follicles obtained after stimulation, reflects the proportion of the variability of the number of follicles explained by this variable(s).
The outcome measure of the second part of this study was the result of ovarian hyperstimulation expressed as the number of retrieved oocytes.
We defined a 'poor' ovarian response as less than 6 oocytes after ovarian hyperstimulation in an IVF treatment and a 'hyper' response as more than 20 oocytes after such an IVF treatment. Among women undergoing in vitro fertilization, the chances of a live birth are related to the number of eggs fertilized, presumably because of the greater selection of embryos for transfer. The low success rate when only two eggs were fertilized reflects the lack of choice among embryos for transfer (Templeton *et al.*, 1998) . We have in our laboratory the experience that we have an overall 50-60 % chance of fertilisation. Taken this togheter, at least 6 oocytes are required for three or more fertilized eggs.
We defined a hyper response when there were > 20 oocytes. This was based on the knowledge that the pregnancy rates do not increase when > 20 oocytes are retrieved. Moreover, such cases have a significant risk of a severe OHSS (Kwee *et al.*, 2006).
In our former study (Kwee *et al.*, 2006), we examined the value of the independent variables by univariate logistic regression: age, bFSH, binhibin B, CCCT-results, E2-increment in EFORT, inhibin B-increment in EFORT. In this study we examined by univariate logistic regression, the value of the independent variables: total basal ovarian volume and the total basal antral follicle count in predicting the ovarian response in predicting a poor and hyper response after ovarian hyperstimulation in IVF. Subsequently multivariate logistic regression analyses were used to develop prediction models for the ovarian response. The area under the

receiver operating characteristic curve (ROC-AUC) was computed to assess the predictive accuracy of the logistic models. To define a 'normal' and an 'abnormal' test, sensitivity, specificity, positive predictive value and accuracy were used to find the optimal cut off level.
Comparison of means was done with the unpaired t-test. For all tests the significance level was 0.05. Statistical analysis of the data was performed with SPSS (Statistical package for Social Sciences; SPSS, Inc., Chicago, IL) for Windows.

Results

The characteristics of the two groups are given as means ± SD in Table 1 (Kwee *et al.*, 2003). No significant differences were noted between the groups in baseline characteristics, cycle day 3 measurements or outcome parameters. In the first group 68 % had a primary infertility and 32 % a secondary infertility. The cause of infertility was for 65 % an idiopathic factor, 31 % a male factor and 4 % a cervical factor. In the second group, 60 % had a primary infertility and 40 % a secondary infertility. The cause of infertility was for 66 % an idiopathic factor, 38 % a male factor and 6 % a cervical factor.
In the CCCT group 32 patients had a normal response to ovarian stimulation, 15 patients had a poor response and 8 patients had a high response. One patient was excluded from analysis because of a severe risk on ovarian hyperstimulation syndrome (OHSS)(the E2 level exeeded 35.000 pmol/l after ovarian hyperstimulation). In the EFORT group 32 patients had a normal response to ovarian stimulation, 14 patients had a poor response and 8 patients had a high response.

Table 1 Characteristics of the groups (values are means ± SD). No significant differences (Kwee *et al,* 2003).

	CCT-group N = 56	EFORT-group N = 54
Baseline characteristics		
Age (y)	33.79 ± 3.95	34.19 ± 3.75
Duration infertility (y)	3.71 ± 2.08	3.87 ± 1,56
Cycle day 3		
FSH (IU/l)	7.60 ± 2.46	7.38 ± 3.11
E2 (pmol/l)	126.05 ± 53.10	118.60 ± 47.06
Inhibin B (ng/l)	94.95 ± 39.36	96.33 ± 40.60
Total volume (ml)	10.91 ± 5.19	12.10 ± 4.69
Total antral follicle count	9.41 ± 5.00	10.66 ± 5.21
Treatment results		
Duration of stimulation (d)	12.4 ± 2.7	11.9 ± 2.3
Number of ampoules of FSH	34.2 ± 8.0	32.7 ± 7.0
E2 level on the day of hCG (pmol/l)	11155.41 ± 18591.13	12134.78 ± 17872.12
Endpoints		
Total number of follicles	14.27 ± 10.23	14.17 ± 10.27
Total number of oocytes	11.58 ± 8.51	11.93 ± 9.11

Univariate linear regression analysis

The correlation between the total basal ovarian volume (BOV) and number of follicles obtained after stimulation and the correlation between the count of the total basal antral follicle (AFC) and number of follicles obtained after stimulation are calculated. The regression line of the basal ovarian volume on the number of follicles (Figure 1A) was drawn by the regression equitation: X = -0.211 + 1.239 x tot. Volume; with a 95% CI of 0.909-1.569, meaning that each increment of 1 ml of ovarian volume predicts an increment of 1.2 follicle (95% CI: 0.9-1.6) (r =0.610, P < 0.001). The regression equitation for the total basal antral follicle (Figure 1B), X = -0.568 + 1.479 x tot. antral foll. (1.222-1.736), shows that an increase of 1 antral follicle predicts an increment of 1.5 follicles (r = 0.741, P < 0.001).

Table 2 shows the results of the EFORT and CCCT as described in our previous study (Kwee *et al.*, 2003) and the additional results of the transvaginal ultrasound.

Figure I. **(A)** Plot of the number of follicles obtained after stimulation against the basal total ovarian volume. The three lines represent the regression line: Y= -0.460 + 1.255 x basal tot. ovarian volume with the 95% confidence interval (CI) of the mean.

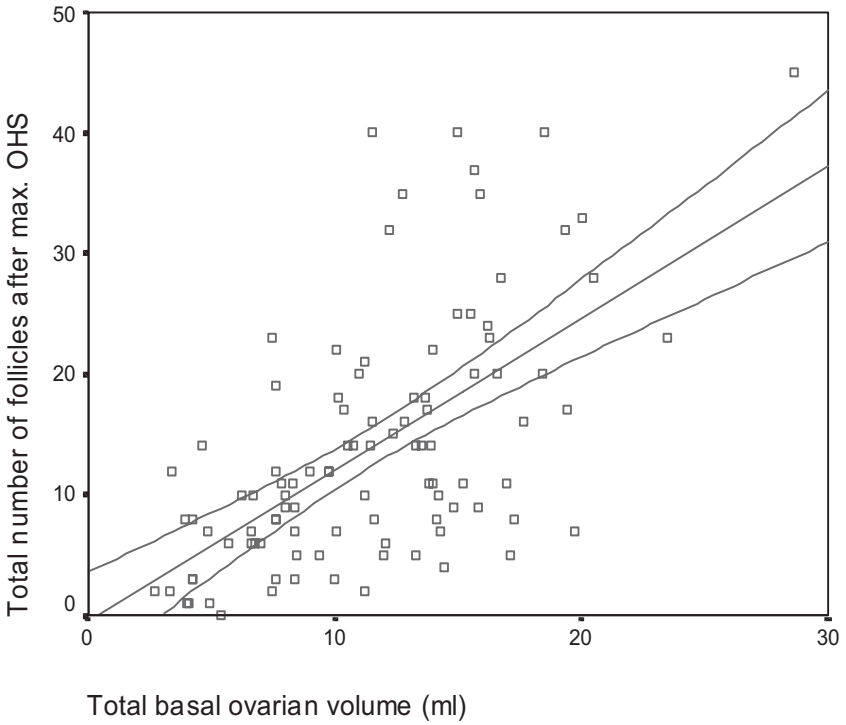


Figure I. (B) Plot of the number of follicles obtained after stimulation against the basal total antral follicle count. The three lines represent the regression line: Y= -0.730 + 1.491 x tot. antral follicle count with the 95% CI of the mean.

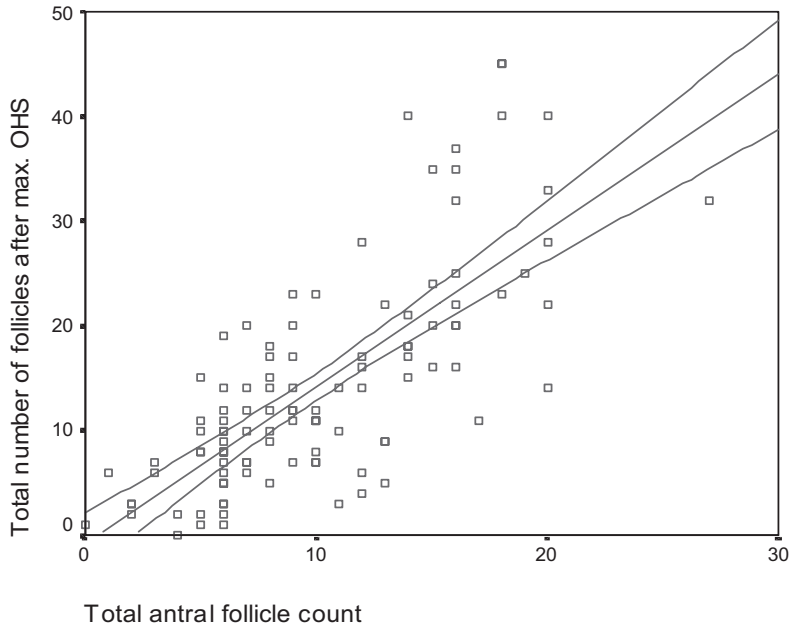


Table 2 Univariate regression analysis of the ovarian reserve tests for the prediction of the stimulative cohort of the ovaries (ovarian reserve) (Kwee *et al.*, 2003)

	N	Correlation	P
Age (y)	110	0.423	< 0.001
bFSH (IU/l)	110	0.313	0.001
∑ bFSH + sFSH in the CCCT (IU/l)	56	0.496	< 0.001
E2-increment in the EFORT (pmol/l)	54	0.751	< 0.001
Inh.B-increment in the EFORT (ng/l)	54	0.718	< 0.001
Total ovarian volume (ml)	110	0.610	< 0.001
Total antral follicle count	110	0.745	< 0.001

Stepforward regression analysis: Prediction model for ovarian reserve

Based on the CCCT group, the prediction model for ovarian response is explained for 51 % by the best predictive variable: the total antral follicle count. When adding the independent variables: total basal volume, ∑ bFSH + sFSH, bFSH and age in a stepforward regression analysis, the explained variation rose significantly with 5 % after the selection of bFSH. The independent variable total basal volume, ∑ bFSH + sFSH and age, did not have a significant

contribution to the model. The exact prediction of the total number of follicles obtained after stimulation thus increased from 51 % to 56 %. The regression line of the bFSH and total antral follicle count on the number of follicles was drawn by the regression equation: Y = 9.478 – 0.985 x bFSH (-1.857- -1.150) + 1.122 x AFC (0.698 -1.561) (r=0.748, p<0.001). Based on the EFORT group, the prediction model for ovarian response is explained for 63 % by the best predictive variable, the total antral follicle count. When adding the Inhibin B-increment and total basal volume simultaneously in a stepforward multiple regression prediction model, the explained variation of the best predictive model rose significantly with 9 %. The total explained variation thus increased from 63 % to 72 %. The regression line of the total antral follicle count, Inhibin B-increment and total basal volume on the number of follicles was drawn by the regression equation: Y = -3.161 + 0.805 x AFC (0.258-1.352) + 0.034 x Inh. B-incr.(0.007-0.601) + 0.511 BOV (0.480-0.974)(r=0.848, p<0.001). When we included E2-increment, age and bFSH as variables in the stepforward regression analysis together with total antral follicle count, Inhibin B-increment and the total basal ovarian volume we did not find a significant contribution of these variables.

Univariate logistic regression

Table 3 depicts the ROC-AUC for the total basal ovarian volume and the total basal antral follicle count for the prediction of poor response after IVF with ovarian hyperstimulation and also the results of the EFORT and CCCT as described previously (Kwee *et al.*, 2006). Both tests have the potential to predict poor response, expressed by the ROC-AUC (0.83 respectively, 0.77).

Table 3 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC-AUC) of the ovarian reserve tests for the prediction of 'poor' respone in IVF (Kwee *et al.*, 2006).

Variable	N	P	ROC AUC
Univariate analysis			
Age (y)	110	0.033	0.63
bFSH (IU/l)	110	< 0.0001	0.83
∑ bFSH + sFSH in the CCCT (IU/l)	56	< 0.0001	0.88
E2-increment in the EFORT (pmol/l)	54	0.006	0.75
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.86
Total ovarian volume (ml)	110	< 0.0001	0.77
Total antral follicle count	110	< 0.0001	0.83
Multivariate analysis			
CCCT GROUP			
∑ bFSH + sFSH in the CCCT (IU/l))	56	< 0.0001	0.88
Multivariate analysis			
EFORT GROUP			
Total antral follicle count	54	0.003	0.88

Table 4 presents test characteristics such as sensitivity, specificity, positive predictive value and accuracy at different cut off levels of the AFC to define a normal (non-poor response) and an abnormal (poor response) test for the prediction of 'poor'response after IVF. The cut off level of < 6 antral follicles had a sensitivity of 41 % and a specificity of 95 %. In the population studied, with a prevalence of 27 % for a poor response (< 6 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 89 % (which means that 89 % of the patients had a correctly predicted test). In case of a result less than 6 antral follicles, the test correctly predicted poor response to stimulation in an IVF-treatment in 75 % (positive predictive value).

Table 4 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for the total antral follicle count (AFC) for the prediction of 'poor'response in IVF.

Total AFC	Sensitivity	Specificity	PPV	Accuracy
< 4	0.21	0.99	0.86	0.78
< 5	0.28	0.99	0.89	0.80
< 6	**0.41**	**0.95**	**0.75**	**0.89**
< 7	0.69	0.80	0.56	0.77
< 8	0.76	0.74	0.51	0.75

Table 5 depicts ROC-AUC for the total basal ovarian volume and the total basal antral follicle count for the prediction of hyper response after IVF with ovarian hyperstimulation and also the results of the EFORT and CCCT as described previously (Kwee *et al.*, 2006). As a single prognostic predictor, the AFC appeared to have a good discriminative potential for hyper response, expressed by a large ROC-AUC (0.92).
Table 6 presents test characteristics such as sensitivity, specificity, positive predictive value and accuracy at different cut off levels of the AFC to define a normal (non-high response) and an abnormal (hyper response) test for the prediction of hyper response after IVF. The cut off level of > 14 antral follicles gave the highest sum of the sensitivity, specificity and gave also the highest accuracy. This result had a sensitivity of 82 % and a specificity of 89%. In the population studied, with a prevalence of 15 % for high response (> 20 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 88 % (which means that 88 % of the patients had a correct predicted test). In case of a result of greater than 14 antral follicles, the test correctly predicted hyper response to stimulation in an IVF-treatment in 58 % (positive predictive value).

Table 5 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC AUC) of the ovarian reserve tests for the prediction of hyper 'response' in IVF (Kwee *et al.*, 2006).

Variable	N	P	ROC AUC
Univariate analysis			
Age (y)	110	0.004	0.71
BFSH (IU/l)	110	< 0.0001	0.80
∑ bFSH + sFSH in the CCCT (IU/l)	56	0.003	0.82
E2-increment in the EFORT (pmol/l)	54	0.003	0.83
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.92
Total ovarian volume (ml)	110	< 0.0001	0.87
Total antral follicle count	110	< 0.0001	0.92
Multivariate analysis			
CCCT GROUP			
Age	56	0.032	} 0.93
Total antral follicle count	56	<0.0001	
Multivariate analysis			
EFORT GROUP			
Total antral follicle count	54	<0.0001	0.93

Table 6 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for the total antral follicle count (AFC) for the prediction of 'hyper'response in IVF.

Total AFC	Sensitivity	Specificity	PPV	Accuracy
> 10	0.94	0.71	0.36	0.76
> 12	0.88	0.80	0.44	0.81
> 14	0.82	0.89	0.58	0.88
> 16	0.47	0.96	0.67	0.88
> 18	0.29	0.98	0.71	0.87

Multivariate logistic regression

In the CCCT group, multivariate analysis for poor response resulted in a model with 1 variable: ∑ bFSH + sFSH in the CCCT (ROC-AUC = 0.88).

In the EFORT group, multivariate analysis for poor response resulted in a model with only one variable: Total antral follicle count (ROC-AUC = 0.88) (Table 3).

In the CCCT group, multivariate analysis for hyper response resulted in a model with 2 variables: age and AFC (ROC-AUC = 0.93).

In the EFORT group, multivariate analysis for hyper response resulted in a model with only one variable: AFC (ROC-AUC = 0.93) (Table5).

Discussion

AFC is able to accurately predict the number of follicles obtained during maximal ovarian stimulation. According to our study that uniquely allowed direct comparison, AFC does not seem superior to other common basal and stimulated endocrine ovarian reserve tests. Included into the stepwise forward multiple regression model it lead, in combination with the Inhibin B-increment in the EFORT and BOV, to the most optimal prediction model. On the other hand, according to logistic analysis, AFC sofar seems to be the only test able to reliably predict low and high responders.

The performance of AFC with regard to the prediction of poor response gave a sensitivity of 73 % and a specificity of 95 %, which would imply that the test performs only moderately, especially at the sensitivity level. In comparison with CCCT this sensitivity is lower whereas specificity seems the same. Consequently there would be more false negative patients with potential undertreatment as result. Increasing the threshold of AFC implies better sensitivity and to some extend a still acceptable specificity. However the decrease of accuracy indicates that overall an unacceptable number of patients will be misdiagnosed.

Recently Hendriks *et al.* (2005b), published a meta-analysis on the AFC as a predictor for poor ovarian response and concluded that AFC is an adequate test for the prediction of poor ovarian response, comparing to bFSH. The data of our study that does meet the criteria for inclusion in this meta analysis would fit seamlessly into the summary ROC curve of report. We confirm the previous observation of the little difference between CCCT and AFC (Hendriks *et al.*, 2005a).

The high intercycle stability of AFC (Bancsi *et al.,* 2004) and its potentially likely attractive cost features, although formal cost effect comparison studies need to be done, are likely to make this test rather attractive for routine practice.

A great advantage of AFC over any other test is its potential usefulness for its ability to concomitantly predict low and high responders. So far EFORT seemed to have the best grades (Kwee *et al.*, 2006) but the current analysis provides evidence that AFC is superior. The test characteristics show us that an AFC > 14 could lead to the decision to adjust the gonadotrophin dose in trying to prevent a hyper response leading to OHSS. Of course the choice of the cut-off level depends on the appreciation of false positive and false negative results and on the consequences drawn by the clinician from an abnormal test.

Total volume of the ovaries detected by transvaginal ultrasound is correlated with the outcome parameters but not better than the count of antral follicles. Its performance was slightly to moderately less than that of AFC, both for poor and high response. Our data agree well with that published in a recent meta analysis (Broekmans *et al.*, 2006).

In conclusion AFC performs well as a test for ovarian response being superior or at least similar to complex expensive and time consuming endocrine tests, probably most applicable in general practise.

Future studies will have to be carried out to determine if other ovarian reserve tests such as the measurement of Anti- Müllerian Hormone (AMH) (Hazout *et al.*, 2004, Seifer *et al.*, 2002, Vet *et al.*, 2002) are better predictors for ovarian reserve.

References

Andolf E, Jörgensen C, Svalenius E and Sundén B (1987) Ultrasound measurement of the ovarian volume. Acta Obstet Gynecol Scand 66, 387-89.

Bancsi LF, Broekmans FJ, Looman CW, Habbema JD and te Velde E.R (2004) Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. Fertil Steril 81, 35-41.

Block E (1952) Quantitative morphological investigations of the follicular system in women. Variations at different ages. Acta Anat 14, 108-23.

Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB (2006) A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod update, in press.

Campbell S, Goessens L, Goswamy R and Whitehead M (1982) Real-time ultrasonography for determination of ovarian morphology and volume. A possible early screening test for ovarian cancer? Lancet i, 425-6.

Chang MY, Chiang CH, Hsieh TT, Soong YK and Hsu KH (1998) Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil. Steril 69, 505-10.

Faddy MJ and Gosden RG (1996) A model conforming the decline in follicle numbers to the age of menopause in women. Hum Reprod 11, 1484-6.
Gougeon A, Ecochard R and Thalabard JC (1994) Age-related changes of the population of human ovarian follicles: increase in the disappearance rate of non-growing and early-growing follicles in aging women. Biol Reprod 50, 653-63.

Gougeon A (1984) Caracteres qualitatifs et quantitatifs de la population folliculaire dans lóvaire humaine adulte. Contracept Fertil Sex 12, 527-35.

Guraya SS (1970) Correlation between the findings of light and electronmicroscopy in human primordial follicles. Acta Anat 77, 617-35.

Hazout A, Bouchard P, Seifer DB, Aussage P, Junca AM and Cohen-Bacrie P (2004) Serum antimullerian hormone/mullerian-inhibiting substance appears to be a more discriminatory marker of assisted reproductive technology outcome than follicle-stimulating hormone, inhibin B, or estradiol. Fertil Steril 82, 1323-9.

Hendriks DJ, Broekmans FJ, Bancsi LF, de Jong FH, Looman CW, Te Velde ER (2005a) Repeated clomiphene citrate challenge testing in the prediction of outcome in IVF: a comparison with basal markers for ovarian reserve. Hum Reprod 20, 163-9.

Hendriks DJ, Mol BWJM, Bancsi LFJMM, te Velde ER and Broekmans FJM (2005b) Antral follicle count in the prediction of poor ovarian response and pregnancy after in-vitro-fertilization: a meta-analysis and comparison with basal follicle-stimulating hormone level. Fertil Steril 83, 291-301.

Higgins RV, van Nagell JR, Woods CH, Thompson EA and Kryscio RJ (1990) Interobserver variation in ovarian measurements using transvaginal sonography. Gynecol Oncol 39, 69-71.

Ivarsson SA, Nillson KO and Persson PH (1983) Ultrasonography of the pelvic organs in prepubertal and postpubertal girls. Arch Dis Child 58, 352-4.

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB and Schoemaker J (2003) Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study. Hum Reprod 18, 1422-7.

Kwee J, Schats R, McDonnell J, Schoemaker J and Lambalk CB (2006) The Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of low and hyperresponders to in vitro fertilization (IVF). Fertil Steril, in press.

Lass A, Skull J, McVeigh E, Margara R and Winston RML (1997) Measurement of ovarian volume by transvaginal sonography before ovulation induction with human menopausal gonadotrophin for in-vitro fertilization can predict poor response. Hum Reprod 12, 294-7.

Latin-American Puregon IVF Study Group (2001) A double-blind clinical trial comparing a fixed daily dose of 150 and 250 IU of recombinant follicle-stimulating hormone in women undergoing in vitro fertilization. Fertil Steril 76, 950-6.

Out HJ, Braat DM, Lintsen BME, Gurgan T, Bukulmez O, Gokmen O, Keles G, Caballero P, Gonzalez JM, Fabregues F et al. (2000) Increasing the daily dose of recombinant follicle stimulating hormone (Puregon®) does not compensate for the age-related decline in retrievable oocytes after ovarian stimulation. Hum Reprod 15, 29-35.

Out HJ, David I, Ron-El R, Friedler S, Shalev E, Geslevich J, Dor J, Shulman A, Ben Rafael Z, Fisch B et al. A randomized, double-blind clinical trial using fixed daily dose of 100 or 200 IU of recombinant FSH in ICSI cycles. Hum Reprod 2001;16:1104-9.

Kline J, Kinney A, Kelly A, Reuss ML, Levin B. Predictors of antral follicle count during the reproductive years. Hum Reprod 2005;20:2179-89.

Richardson SJ, Senikas V, Nelson JF. Follicular depletion during the menopausal transition: evidence for accelerated loss and ultimate exhaustion. J Clin Endocrinol Metab 1987;65:1231-7.

Scheffer GJ, Broekmans FJ, Dorland M, Habbema JD, Looman CW, te Velde ER. Antral follicle counts by transvaginal ultrasonography are related to age in women with proven natural fertility. Fertil Steril 1999;72: 845–51.

Scheffer GJ, Broekmans FJM, Looman CWN, Blankenstein M, Fauser BCJM, de Jong FH, te Velde ER. The number of antral follicles in normal women with proven fertility is the best reflection of reproductive age. Hum Reprod 2003;18:700-6.

Seifer DB, Mac Laughlin DT, Christian BP, Feng B, Shelden RM. Early follicular serum mullerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. Fertil Steril 2002;77:468-71.

Syrop CH, Willhoite A, van Voorhis BJ. Ovarian volume: a novel outcome predictor for assisted reproduction. Fert Steril 1995;64:1167-71.

Templeton A and Morris JK. Reducing the risk of multiple births by transfer of two embryos after in vitro fertilization. N Eng J Med 1998;339:573-7.

Thatcher SS and Naftolin F. The aging and aged ovary. Semin Reprod Endocrinol 1991;9:189-99.

Tomás C, Nuojua-Huttunen S, Martikainen H. Pretreatment transvaginal ultrasound examination predicts ovarian responsiveness to gonadotrophins in in-vitro fertilization. Hum Reprod 1997;12:220-3.

Van der Meer M, Hompes PGA, de Boer JAM, Schats R, Schoemaker J. Cohort size rather than follicle-stimulating hormone threshold level determines ovarian sensitivity in polycystic ovary syndrome. J Clin Endocrinol Metab 1998;83:423-26.

Vet A, Laven JSE, de Jong FH, Themmen APN, Fauser BCJM. Antimüllerian hormone serum levels: a putative marker for ovarian aging. Fertil Steril 2002;77:357-62.

Chapter 6

Evaluation of AMH as test for the prediction of ovarian reserve

J. Kwee[1], R. Schats[1],J. McDonnell[1], A.P.N. Themmen[2], F.H. de Jong[2], and C.B. Lambalk[1]

[1] Division of Reproductive Endocrinology and Fertility and the IVF Centre, department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, the Netherlands.
[2] Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands.

Human Reproduction 2003;18:1422-7

Abstract

Study Objective: This study aims to compare in an integral way the value of the serum basal Anti-Müllerian hormone (AMH) level with most of the established ovarian reserve tests.
Design: Prospective randomized controlled trial
Setting: Fertility centre of an university hospital
Patients: 110 patients undergoing their first IVF cycle, randomized, by a computer designed 4-blocks system study into two groups.
Interventions: Fifty six patients underwent a Clomiphene Citrate Challenge Test (CCCT), and 54 patients underwent an Exogenous FSH Ovarian Reserve Test (EFORT). In all patients basal AMH, basal FSH, basal inhibin B, antral follicle count (AFC) and basal volume of the ovaries (BOV) were measured. In all patients, the test was followed by a standard IVF treatment.
Main outcome measure(s): Ovarian response after ovarian hyperstimulation in an IVF treatment, expressed as the total number of stimulated follicles, retrieved oocytes and ongoing pregnancies.
Results(s): The best prediction of ovarian reserve (Y) was seen in a multiple regression prediction model that included, AFC, Inhibin B-increment in the EFORT and BOV simultaneously (Y = -3.161 + 0.805 x AFC (0.258-1.352) + 0.034 x Inh. B-incr.(0.007-0.601) + 0.511 BOV (0.480-0.974)($r=0.848$, $p<0.001$). Univariate logistic regression showed that the best predictors for poor response were AMH (ROC-AUC = 0.85), the CCCT (ROC-AUC = 0.87), bFSH (ROC-AUC = 0.83) and the AFC (ROC-AUC = 0.83). Multiple logistic regression analysis did not produce a better model in terms of improving the prediction of poor response. For hyper response, univariate logistic regression showed that the best predictors were AFC (ROC-AUC = 0.92) and the inhibin B-increment in the EFORT (ROC-AUC = 0.92), but AFC had better test characteristics, namely a sensitivity of 82 % and a specificity of 89 %. Multiple logistic regression analysis did not produce a better model in terms of predicting hyper response. The best predictors for the prediction of non-pregnancy were the CCCT (ROC-AUC = 0.75) and the E2-increment in the EFORT (ROC-AUC = 0.71).
Conclusion(s): AMH is comparable with other commonly used ovarian reserve tests, but is probably most applicable in general practice, because it can be measured throughout the cycle.

Key Words: AMH/bFSH/bInhibin B/CCCT/EFORT/IVF/ovarian reserve

INTRODUCTION

Will I become pregnant: yes or no, that is what couples, who are seeking help in an infertility clinic, want to hear about their chances. The age-related decline of the success in IVF is largely attributable to a progressive decline of ovarian oocyte quality and quantity (Broekmans *et al*., 2006). Over the past two decades, a number of so-called ovarian reserve tests (ORTs) have been designed to give an answer to that particular question. A potential new test in this field is measuring levels serum Anti-Müllerian hormone (AMH), a dimeric glycoprotein, which is a member of the Transforming Growth Factors-β (TGF-β) family (Van Rooij *et al*., 2002, Van Rooij *et al*., 2004, Van Rooij *et al*., 2005, Fanchin *et al*., 2003, Seifer *et al*., 2002). After follicles differentiate from the primordial to the primary stage, production of AMH starts and it continues until the follicles have reached the antral stages (Durlinger *et al*., 1999, Durlinger *et al*., 2002, Weenen *et al*., 2004). The number of the small antral follicles is related to the size of the primordial follicle pool (Kevenaar *et al*., 2006).
Recently, serum AMH has been shown to be a marker for the size of the residual follicle pool and its decline follows the ageing process in a more gradual fashion (Van Rooij *et al*., 2002, Van Rooij *et al*., 2004, Van Rooij *et al*., 2005, Fanchin *et al*., 2003, Seifer *et al*., 2002) the currently used ovarian reserve tests, so it is worthwhile to evaluate the value of serum AMH as a marker of ovarian reserve.
A few studies tested the correlation between serum AMH and the number of oocytes after ovarian hyperstimultion in an IVF procedure. They all found that serum AMH may serve as a good candidate for the determination of the ovarian reserve (Van Rooij *et al*., 2002, Fanchin *et al*., 2003, Seifer *et al*., 2002). Moreover, evidence is accumulating that serum AMH, in contrast to FSH, estradiol (E2) and inhibin B, can be used as a cycle independent marker (Hehenkamp *et al*., 2006, La marca *et al*., 2006).
Recently we conducted a prospective study that evaluated various established static and dynamic ovarian reserve test (Kwee *et al*., 2003, Kwee *et al*., 2006, Kwee *et al*., submitted) but serum AMH was not included. Given its potential features as a reliable test we decided to retrospectively measure serum AMH in remaining blood samples, allowing us to make an integral comparison of serum AMH levels as a test to predict ovarian response to gonadotropin stimulation with most of the established ovarian reserve tests including basal FSH, basal inhibin B, Clomiphene Citrate Challenge test (CCCT), Exogenous FSH Ovarian Reserve Test (EFORT), antral follicle count (AFC) and basal volume of the ovaries (BOV).

MATERIALS AND METHODS

Study Population

One hundred and ten patients, aged 18-39 years, who were eligible for treatment by assisted reproduction between June 1997 and December 1999 participated in the study. This study is part of a prospective randomized study on the determination of ovarian reserve in regular menstruating patients (Kwee *et al*., 2003). Their infertility was either idiopathic for > 3 years and/or due to a male factor and/or cervical hostility. Patients had to have regular menstrual cycles, two ovaries and at least one patent Fallopian tube. Excluded were patients with either polycystic ovary syndrome, defined as a combination of oligo- or amenorrhoea and an increased luteinizing hormone (LH) concentration in the presence of a normal follicle

stimulating hormone (FSH) level or a severe male factor, defined as 1. less than 1 million motile spermatozoa after centrifugation (40/90) and/or 2. > 20 % antisperm antibodies present on the spermatozoa after processing with gradient centrifugation (40/90) and/or 3. > 50 % of the spermatozoa without an acrosome. Other exclusion criteria were untreated or insufficiently corrected endocrinopathies, clinically relevant systemic diseases or a body mass index > 28 kg/m².
The protocol was approved by the Institutional review Board and the Committee on ethics of research involving human subjects of the Vrije Universiteit Medical Centre, Amsterdam, the Netherlands. All couples participating in the study signed informed consent.

Treatment protocol
In our previous studies (Kwee *et al*., 2003, Kwee *et al*., 2006, Kwee *et al*., submitted), 110 patients were randomized by a computer-designed 4-blocks system into two groups. Fifty-six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by IVF treatment. From all blood samples, serum and plasma were separated and stored at – 20°C for later estimation of levels of serum AMH.
For the purpose of this study, we analysed the basal serum AMH (bAMH) on cycle day 3 (CD 3, CD 1 is the day of onset of menses) and additionally on CD 4 in the EFORT and on CD 10 in the CCCT. The bFSH level, bE2 level and bInhibin B level were determined as an integral part of all CCCT's and EFORT's, as described previously (14). On cycle day 3, all patients underwent a transvaginal ultrasound examination to assess the number of antral follicles and the volume of the ovaries.

Clomiphene citrate challenge test (CCCT): starting on CD 5 100 mg of Clomiphene citrate (Serophene®; Serono, Geneva, Switzerland) was administered for 5 days. Serum FSH was determined on CD 2 or 3 (bFSH) and on CD 10 (sFSH). Analysis of the CCCT (Kwee *et al*., 2003) was performed using the parameter: bFSH + sFSH.

Exogenous Follicle stimulating hormone Ovarian Reserve Test (EFORT): on CD 3, 300 IU recFSH (Gonal-F®, Serono, Geneva, Switzerland) were administered subcutaneously (s.c). In this study blood samples for the determination of FSH, E2 and Inhibin B were drawn: just before (basal values) and 24 hrs after (stimulated values) the administration of FSH. Analysis of the EFORT (14) included the following parameters: E2-increment and Inhibin B-increment 24 hrs after administration of FSH.

Transvaginal Sonography Measurements
All ultrasound examinations were performed by two the authors (J.K or R.S) using an Aloka SSD-1700 ultrasound apparatus (5.0 MHz probe).
The volume of each ovary was calculated by measuring the ovarian diameters (D) in three perpendicular directions and applying the formula for an ellipsoid: (D1 x D2 x D3 x □/ 6). The volumes of both ovaries were added (to obtain the BOV).
To determine the diameter of a follicle, the mean of measurements in two perpendicular directions was taken. The numbers of follicles in both ovaries were added for the total AFC. The follicles visualized and counted by transvaginal sonography (TVS) in the early follicular phase are 2-10 mm in size.
IVF-treatment: The ovarian hyperstimulation protocol was performed according to a long GnRH-agonist protocol starting in the midluteal phase. On CD 3 of the first cycle the ovarian

volume and antral follicle count were measured by TVS examinations. In the subsequent midluteal phase, seven days after ovulation, daily s.c. injections with triptoreline-acetate (Decapeptyl®, 0.1 mg/day; Ferring, Hoofddorp, the Netherlands) were started. Because of the administration of the GnRH-agonist, patients were advised to use a barrier type of contraception during this cycle.
On CD 3 of the next cycle, ovarian hyperstimulation was started with daily s.c. injections of a fixed dose of 225 IU uFSH (Metrodin HP®, 75 IU/amp; Serono, Geneva, Switzerland).
Van der Meer *et al*. (1998) showed that in eumenorrheic patients, the median (range) FSH threshold level for monofollicular growth was 5.3 (4.3-8.2) IU/l and the median (range) threshold dose was 75 IU (0.5-1.75 IU) FSH/day. It was concluded that by an increment of 37.5 IU of FSH above the threshold dose for monofollicular growth, the maximum response is already obtained. It seems that in IVF stimulation a maximal effect is reached with FSH dosages up to 225 IU (Latin-American Puregon IVF Study Group, 2001, Out *et al*., 2000, Out *et al*., 2001). Combining these facts, it can be concluded that an initial stimulation with 225 IU of FSH under a long (GnRH agonist suppressed) protocol, probably gives a maximal IVF stimulation, the outcome of which could be used as the gold standard for the cohort size.
Standard procedures were followed including TVS as described above on CD 2 or 3 and on CD 9 or 10. Daily TVS was performed from the moment when the leading follicle reached a diameter of 16 mm. Ovarian hyperstimulation was continued until the largest follicle reached a diameter of at least 18 mm. The maximum duration of uFSH administration allowed was 16 days. If these criteria were met, Metrodin HP® and Decapeptyl® were discontinued and 10.000 IU of hCG (Profasi®, 10.000 IU/amp; Serono, Geneva, Switzerland) was administered. On the day of hCG, TVS was performed to count the result of ovarian hyperstimulation (all follicles ≥ 10 mm) expressed as the total number of follicles. TVS guided follicular aspiration (FA) was performed 36 hours after hCG administration. On the day of hCG administration E2 was determined. Follicular aspiration was done under systemic analgesia (7.5 mg dormicum orally and 50-100 mg pethidine hydrochloride intramuscularly), and all follicles present were aspirated. A maximum of two embryos was transferred. To support the luteal phase micronized progesterone (Progestan®; Nourypharma BV, Oss, the Netherlands) was used.

Ongoing pregnancy
Ongoing pregnancy was defined as the presence of fetal cardiac activity beyond 12 weeks of gestation. For this study, a multiple pregnancy was regarded as one pregnancy.

Serum assays
Serum E2 was determined by a competitive imunoassay (Amerlite, Amersham, UK). For E2, the inter-assay CV was 11 % at 250 pmol/l and 8 % at 8000 pmol/l, the intra-assay coefficient of variation (CV) was 10 % at 350 pmol/l. 8 % at 1100 pmol/l and 8 % at 5000 pmol/l. The lower limit of detection for E2 was 90 pmol/l. In the EFORT and CCCT we measured E2 by a sensitive radioimmunoassay (Sorin, Biomedica, Saluggia, Italy). This measurement of E2 was abbreviated as EE. For EE, the inter-assay CV was 11 % at 60 pmol/l, 8 % at 200 pmol/l, 11 % at 550 pmol/l and 8 % at 900 pmol/l. The intra-assay CV was 4 % at 110 pmol/l and 5 % at 1000 pmol/l. The lower limit of detection for EE was 18 pmol/l. FSH was determined by a commercially available immunometric assay (Amerlite, Amersham, UK). For FSH, the inter-assay CV was 9 % at 3 IU/l and 5 % at 35 IU/l, the intra-assay CV was 9 % at 5 IU/l, 8 % at 15 IU/l and 6 % at 40 IU/l. The lower limit of detection for FSH was 0.5 IU/l. Inhibin

B was determined immunometrically by a commercially available assay (Serotec Limited Oxford UK). For Inhibin B, the inter-assay CV was 17 % at 25 ng/L, 14 % at 55 ng/L and 9 % at 120 ng/L and the intra-assay CV was 8 % till 40 ng/l and 5 % at > 40 ng/l. The lower limit of detection for Inhibin B was 13 ng/l.
Half-way through the study (after 62 patients), the Amerlite assay used to assess FSH was withdrawn from the market and was replaced by another commercially available assay (Delfia, Wallac, Finland). The two assays have been compared and showed excellent linear correlation, although a shift in the values took place (Delfia FSH = 1.28 x Amerlite FSH + 0.01 (r=0.9964)). For the Delfia FSH, the inter-assay CV was 5 % at 3.5 IU/l and 3 % at 15 IU/l. All FSH determinations have been recalculated and are expressed according to the Delfia assay. The lower limit of detection for FSH was 0.5 IU/l.
Values below the detection limit of an assay were assigned a value equal to the detection limit of that assay.
Serum AMH levels were estimated using an enzyme-immunometric assay (Diagnostic Systems Laboratories, Webster, TX). Inter- and intraassay coefficients of variation (CVs) were less than 7%. The detection limit of the assay was 0.026 µg/l. Repeated freezing and thawing of the samples or storage at 37 °C for 1 h did not affect results of the assay.

Statistical analysis
The outcome measure of the first part of this study was the result of ovarian hyperstimulation expressed as the number of follicles. By univariate linear regression, we estimated the value of the independent variable: bAMH. Subsequent multivariate linear regression analysis was used to develop prediction models for the ovarian response.
The outcome measure of the second part of this study was the result of ovarian hyperstimulation expressed as the number of retrieved oocytes and an ongoing pregnancy. As described previously (Kwee *et al.*, 2006), we arbitrarily defined a 'poor' ovarian response as less than 6 oocytes after ovarian hyperstimulation in an IVF treatment and a 'hyper' response as more than 20 oocytes after such an IVF treatment. In the analysis of 'poor' ovarian response, patients with cancelled cycles due to an exaggerated response were included in the group of 'normal' responders. For the analysis of 'hyper' ovarian response patients with cancelled cycles due to an exaggerated response were included in the group of 'hyper' responders.
By univariate logistic regression, we examined the value of the independent variable: bAMH in predicting poor and hyper response and the presence of an ongoing pregnancy after ovarian hyperstimulation in IVF. Subsequent multivariate logistic regression analysis was used to develop prediction models for the ovarian response. The area under the receiver operating characteristic curve (ROC-AUC) was computed to assess the predictive accuracy of the logistic models.
To define a 'normal' and an 'abnormal' test, sensitivity, specificity, positive predictive value and accuracy were used to find the optimal cut off level.
Comparison of means was done with the unpaired t-test and the and Kruskall-Wallis test. For all tests the significance level was 0.05.
Statistical analysis of the data was performed with SPSS (Statistical package for Social Sciences; SPSS, Inc., Chicago, IL) for Windows.

RESULTS

The characteristics of the 2 groups are given as means ± SD in Table 1 (Kwee *et al*., 2003). No significant differences were noted between the groups in baseline characteristics, cycle day 3 measurements or outcome parameters.
There were no significant changes in serum levels of AMH after injection of the FSH in the EFORT and after 5 days of daily 100 mg clomiphene citrate (table 2).
In 6 patients we could not measure serum AMH, because samples were not available anymore.

Table 1 Characteristics of the groups (values are means ± SD). No significant differences

	CCCT-group N = 56	EFORT-group N = 54
Baseline characteristics		
Age (y)	33.8 ± 4.0	34.2 ± 3.8
Duration infertility (y)	3.7 ± 2.1	3.9 ± 1.6
Primary infertility	57.1 %	65 %
Secondary infertility	42.9 %	35 %
Cause of infertility		
- idiopathic factor	35 (62.5 %)	30 (55.5 %)
- male factor	16 (28.6 %)	23 (42.5 %)
- cervical factor	5 (8.9 %)	1 (2 %)
Cycle day 3		
FSH (IU/l)	7.6 ± 2.5	7.4 ± 3.1
E2 (pmol/l)	126.1 ± 53.1	118.6 ± 47.1
Inhibin B (ng/l)	95.0 ± 39.4	96.3 ± 40.6
AMH (µg/l)	2.6 ± 2.3	3.3 ± 2.9
Treatment results		
Duration of stimulation (d)	12.4 ± 2.7	11.9 ± 2.3
Number of ampoules of FSH	34.2 ± 8.0	32.7 ± 7.0
E2 level on the day of hCG (pmol/l)	11155 ± 18591	12134 ± 17872
Endpoints		
Total number of follicles	14.3 ± 10.2	14.2 ± 10.3
Total number of oocytes	11.6 ± 8.5	11.9 ± 9.1
Adequate ovariele respons	32 (57.2 %)	32 (59.3 %)
Poor ovarian response	15 (26.7%)	14 (25.9 %)
Hyper ovarian response	9 (16.1 %)	8 (14.8)
Ongoing pregnancy	12 (21.1 %)	12 (22.2 %)

Table 2 AMH, E2 and Inhibin B concentrations in CCCT and EFORT

N = 53	CD 3	CD 10 in CCCT	P
FSH (IU/L)	7.6 ± 2.5	8.3 ± 5.3	0.21
AMH (µg/l)	2.6 ± 2.3	3.15 ± 2.71	0.95
E2 (pmol/L)	126.1 ± 53.1	1387.5 ± 782.9	< 0.001
Inhibin B (ng/L)	95.0 ± 39.4	317.0 ± 183.3	< 0.001

Values are presented as median (ranges).

N = 51	CD 3	CD 4 in EFORT	P
FSH (IU/L)	7.4 ± 3.1	12.2 ± 9.2	0.04
AMH (µg/l)	3.3 ± 2.9	2.86 ± 2.39	0.70
E2 (pmol/L)	118.6 ± 47.1	288.6 ± 175.2	< 0.001
Inhibin B (ng/L)	96.3 ± 40.6	211.7 ± 127.4	< 0.001

Values are presented as median (ranges).

The prediction of the number of follicles after stimulation

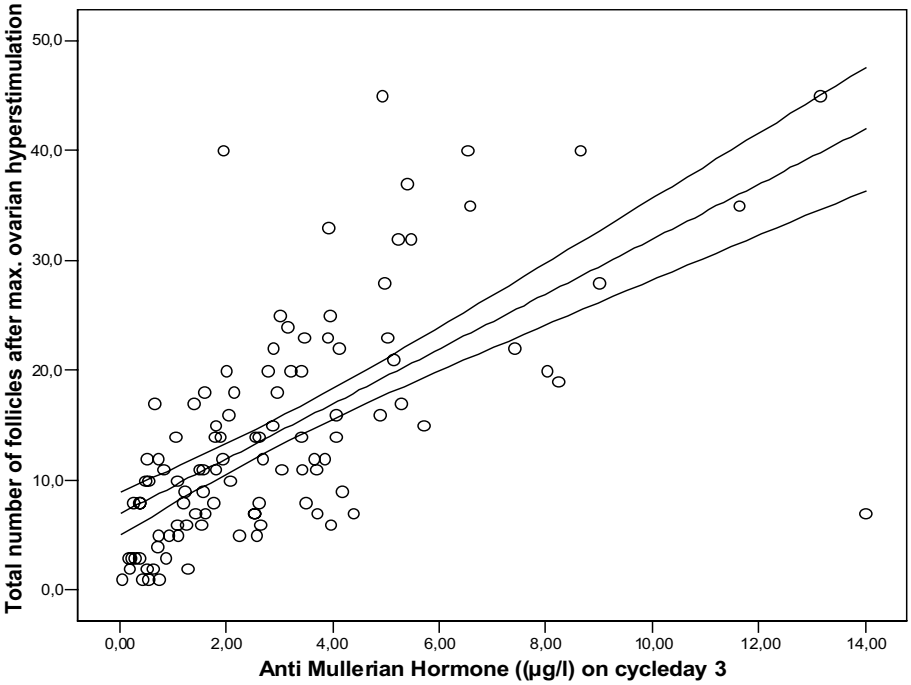
Univariate linear regression analysis

Basal AMH was significantly correlated with the and number of follicles obtained after stimulation (r = 0.632, P < 0.001). The regression line of bAMH versus the number of follicles (Y) was characterized by the equitation: Y = 7.06 + 2.48 x bAMH; with a 95% CI of 1.88 - 3.08, meaning that each increment of 1 µg AMH/l predicts an increment of 2.5 follicle (95% CI: 1.9-3.1) (fig.I). Table 3 shows correlations of numbers of follicles after stimulation with the results of EFORT, CCCT, ultrasound and values of basal estimations as described in our previous study (Kwee *et al*., 2003, Kwee *et al*., 2006, Kwee *et al*., submitted) and the additional results of bAMH measurements.

Table 3 Univariate regression analysis of the ovarian reserve tests for the prediction of the stimulable cohort of the follicles in the ovaries (ovarian reserve).

	N	Correlation	P
Age (y)	110	0.423	< 0.001
bFSH (IU/l)	110	0.313	0.001
Sum of FSH in the CCCT (IU/l)	56	0.496	< 0.001
E2-increment in the EFORT (pmol/l)	54	0.751	< 0.001
Inh.B-increment in the EFORT (ng/l)	54	0.718	< 0.001
Total ovarian volume (ml)	110	0.610	< 0.001
Total antral follicle count	110	0.745	< 0.001
bAMH (µg/l)	104	0.632	< 0.001

Figure I. Plot of the number of follicles obtained after stimulation against the basal AMH. The three lines represent the regression line: Y = 7.06 + 2.48 x bAMH with the 95% confidence interval (CI) of the mean.



Stepforward regression analysis

Based on the CCCT group, the prediction model for ovarian response is explained for 51 % by the best predictive variable: the total antral follicle count. When adding the independent variables: bAMH, total basal volume, bFSH + sFSH, bFSH and age in a stepforward regression analysis, the explained variation rose significantly with 6 % after the selection of bAMH. The other independent variables did not contribute significantly to the model. The prediction of the total number of follicles obtained after stimulation thus increased from 51 % to 57 %. The regression line of the bAMH and total antral follicle count versus the number of follicles (Y) was given by the equation: Y = - 1.052 + 1.019 x bAMH (95% CI: 0.227 - 1.811) + 1.089 x AFC (95% CI: 0.673 -1.504) ($r=0.756$, $p<0.001$).

Based on the EFORT group, the prediction model for ovarian response is explained for 63 % by the best predictive variable, the total antral follicle count. When adding the Inhibin B-increment and total basal volume simultaneously in a stepforward multiple regression prediction model, the explained variation of the best predictive model rose significantly with 9 %. The total explained variation thus increased from 63 % to 72 %. The regression line of the total antral follicle count, Inhibin B-increment and total basal ovarian volume on the number of follicles was drawn by the regression equation: Y = -3.161 + 0.805 x AFC (95% CI: 0.258-1.352) + 0.034 x Inh. B-incr.(95% CI: 0.007-0.601) + 0.511 BOV (95% CI: 0.480-0.974)($r=0.848$, $p<0.001$). When we included bAMH, E2-increment, age and bFSH as variables in the stepforward regression analysis together with total antral follicle

count, Inhibin B-increment and the total basal ovarian volume we did not find a significant contribution of these variables.

The prediction of the number of retrieved oocytes after stimulation

Univariate logistic regression

Table 4 depicts the statistical significance and areas under the receiver operating characteristic curve (ROC-AUC) of logistic regression analysis for bAMH and the other ORT's for the prediction of poor response after IVF with ovarian hyperstimulation.

Table 4 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC-AUC) of the ovarian reserve tests for the prediction of 'poor' response in IVF.

Variable	N	P	ROC AUC
Age (y)	110	0.033	0.63
bFSH (IU/l)	110	< 0.0001	0.83
Sum of FSH in the CCCT (IU/l)	56	< 0.0001	0.88
E2-increment in the EFORT (pmol/l)	54	0.006	0.75
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.86
Total ovarian volume (ml)	110	< 0.0001	0.77
Total antral follicle count	110	< 0.0001	0.83
bAMH (µg/l)	104	< 0.0001	0.85
Multivariate analysis			
CCCT GROUP			
Sum of FSH in the CCCT (IU/l)	56	< 0.0001	0.88
Multivariate analysis			
EFORT GROUP			
Total antral follicle count	54	0.003	0.88

Table 5 presents test characteristics such as sensitivity, specificity, positive predictive value and accuracy at different cut off levels of the bAMH to define a normal (non-poor response) and an abnormal (poor response) test for the prediction of 'poor'response after IVF.

The cut off level of ≤ 1.4 µg/l had a sensitivity of 76 % and a specificity of 86 %. In the population studied, with a prevalence of 27 % for a poor response (< 6 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 83 % (which means that 83 % of the patients had a correctly predicted test). In case of bAMH ≤ 1.4 µg/l, the test correctly predicted poor response to stimulation in an IVF-treatment in 67 % (positive predictive value).

Table 6 depicts the statistical significance and areas under the receiver operating characteristic curve (ROC-AUC) of logistic regression analysis for bAMH and 7 ovarian reserve tests for the prediction of hyper response after IVF with ovarian hyperstimulation. As a single prognostic predictor, the bAMH, appeared to have a good discriminative potential for hyper response, as expressed by a ROC-AUC of 0.85.

Table 5 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for AMH for the prediction of 'poor'response in IVF.

AMH (µg/l)	Sensitivity	Specificity	Positive predictive value	Accuracy
≤ 0.8	0.55	0.94	0.76	0.83
≤ 1.0	0.66	0.94	0.79	0.86
≤ 1.2	0.69	0.88	0.69	0.83
≤ 1.4	**0.76**	**0.86**	**0.67**	**0.83**
≤ 1.6	0.79	0.78	0.58	0.78

Table 6 Univariate and multivariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC AUC) of the ovarian reserve tests for the prediction of hyper 'response' in IVF.

Variable	N	P	ROC AUC
Univariate analysis			
Age (y)	110	0.004	0.71
BFSH (IU/l)	110	< 0.0001	0.80
Sum of FSH in the CCCT (IU/l)	56	0.003	0.82
E2-increment in the EFORT (pmol/l)	54	0.003	0.83
Inh.B-increment in the EFORT (ng/l)	54	< 0.0001	0.92
Total ovarian volume (ml)	110	< 0.0001	0.87
Total antral follicle count	110	< 0.0001	0.92
bAMH (µg/l)	104	< 0.0001	0.85
Multivariate analysis			
CCCT GROUP			
Age	56	0.032	} 0.93
Total antral follicle count	56	< 0.0001	
Multivariate analysis			
EFORT GROUP			
Total antral follicle count	54	< 0.0001	0.93

Table 7 presents test characteristics such as sensitivity, specificity, positive predictive value and accuracy at different cut off levels of the bAMH to define a normal (non-hyper response) and an abnormal (hyper response) test for the prediction of hyper response after IVF. The cut off level of ≥ 5 µg/l gave the highest sum of sensitivity and specificity. This result had a sensitivity of 53 % and a specificity of 91%. In the population studied, with a prevalence of 16 % for high response (> 20 oocytes after ovarian hyperstimulation in an IVF treatment), the accuracy was 85 % (which means that 85 % of the patients had a correctly predicted test). In case of a ≥ 5 µg/l bAMH, the test correctly predicted hyper response to stimulation in an IVF-treatment in 53 % (positive predictive value).

Table 7 Sensitivity, specificity, positive predictive value (PPV) for positive test results and proportion of patients (accuracy) with a correct prediction at different cut off levels for AMH for the prediction of 'hyper'response in IVF.

AMH (µg/l)	Sensitivity	Specificity	PPV	Accuracy
≥ 4.0	0.44	0.84	0.44	0.87
≥ 5.0	0.53	0.91	0.53	0.85
≥ 6.0	0.35	0.96	0.60	0.86
≥ 7.0	0.24	0.96	0.50	0.84
≥ 8.0	0.24	0.97	0.57	0.85

Multivariate logistic regression

In the CCCT group, multivariate analysis for poor response resulted in a model with 1 variable: bFSH + sFSH in the CCCT (ROC-AUC = 0.88).

In the EFORT group, multivariate analysis for poor response resulted in a model with only one variable: AFC (ROC-AUC = 0.88) (Table 3).

In the CCCT group, multivariate analysis for hyper response resulted in a model with 2 variables: age and AFC (ROC-AUC = 0.93).

In the EFORT group, multivariate analysis for hyper response resulted in a model with only one variable: AFC (ROC-AUC = 0.93) (Table5).

The prediction of ongoing pregnancy after IVF treatment

Table 8 depicts the statistical significance and areas under the receiver operating characteristic curve (ROC-AUC) of logistic regression analysis for bAMH and 7 ovarian reserve tests for the prediction of non-pregnancy after IVF with ovarian hyperstimulation. The CCCT with a cut off level of 18 IU/l (sensitivity of 25 % and a specificity of 100 %) and the E2-increment in the EFORT with a cut of level of 130 ng/l (15) (sensitivity of 45 % and a specificity of 83 %) appeared to have the best discriminative potential for the prediction of non-pregnancy. The ROC-AUC of the CCCT is 0.745 with a sensitivity of 0.73 and specificity of 1.0. The ROC-AUC of the E2-increment in the EFORT is 0.709 with a sensitivity and specificity of resp. 0.76 and 0.83.

Table 8 Univariate logistic regression analysis and areas under the receiver operating characteristic curve (ROC-AUC) of the ovarian reserve tests in the prediction of non- pregnancy in IVF.

Variable	N	P	ROC AUC
Age (y)	110	0.108	0.598
BFSH (IU/l)	110	0.087	0.587
Sum of FSH in the CCCT (IU/l)	56	0.005	0.745
E2-increment in the EFORT (pmol/l)	54	0.024	0.709
Inh.B-increment in the EFORT (ng/l)	54	0.119	0.638
Total ovarian volume (ml)	110	0.052	0.629
Total antral follicle count	110	0.080	0.608
bAMH (µg/l)	104	0.023	0.643

DISCUSSION

Serum AMH seems able to predict the number of follicles obtained during maximal ovarian stimulation. According to our study that uniquely allowed direct comparison, bAMH does not seem superior to the antral follicle count, basal ovarian volume and most of the other commonly used stimulated endocrine ovarian reserve tests, providing similar AUC-ROC values. Included into the stepwise forward multiple regression model bAMH did not have additive value to a combination of the Inhibin B-increment in the EFORT and BOV, which led to the most optimal prediction model with regard to ovarian response.

The performance of bAMH with regard to the prediction of poor response gave a sensitivity of 76 % and a specificity of 86 %, which would imply that the test performs moderately. Increasing the threshold of bAMH yields a better sensitivity but a non acceptable specificity and by decreasing the threshold the sensitivity will drop in favour of a higher specificity. In comparison with CCCT this sensitivity and specificity is lower, which means that bAMH has no additional value as a test for poor responders.

As a test to predict ovarian hyper response bAMH does not seem appropriate, the ROC-AUC of the inhibin B-increment in the EFORT and AFC being higher than that for bAMH. The sensitivity is low using all acceptable threshold levels, which means that there would be a lot of false negative patients with potential overtreatment as result.

The results of bAMH in this study confirm the outcome of our recently published systematic review on ovarian reserve tests (Broekmans *et al*., 2006), that bAMH reasonably predicts ovarian response but unfortunately not pregnancies. With regard to bAMH, the data of this study could not be included anymore but would fit seamlessly into the summary ROC curve of the report.

The CCCT and EFORT had a better predictive value for the prediction of pregnancy. An ideal ovarian reserve test should identify a substantial percentage of IVF indicated cases, which have a practically zero chance of becoming pregnant in a series of treatment cycles due to the adverse effects of diminished ovarian reserve. Those cases can be refrained from entering the program, as they will cause very high costs for only minimal results. If not too expensive and not too demanding for the patient, such a test would be readily embraced by physicians, patients, health politicians and insurance companies. In case of the CCCT, there was a specificity of 100%, which means that there were no ongoing pregnancies above the test result of 18 IU/l, but with a very low sensitivity of 25%. It should be noted that the use of pregnancy as outcome parameter for the assessment of ovarian reserve status may be insufficient if only one exposure cycle is taken into account (Broekmans *et al*., 2006). As such, the possibility of misjudgment on the basis of currently known ovarian reserve test is hard to rule out. This implies that the use of the test as a method to deny treatment to assumed ovarian aged women should be declined and, as a consequence the test should not be applied on a regular basis or only used for counseling or screening purposes.

As shown in table 9 poor ovarian response has been associated with a reduced chance of pregnancy in the actual treatment cycle. Accurate prediction of poor response could therefore have clinical value if the pregnancy prospects are so unfavorable that a predicted poor responder would be denied treatment. Accuracy in response prediction, however, will only be high if the false positives are prevented by using extreme cut off levels, implicating that only minor percentages of abnormal tests will be found and many future poor responders will pass unrecognized. At the same time it is necessary to know whether the predicted poor

responder indeed has very low prospects for success in subsequent cycles. As much of this is unknown at the present time, the denial of enetering an IVF treatment on basis of an ovarian reserve test should not be supported.
There are potential advantages of using bAMH over AFC or the CCCT, because AMH can be measured throughout the cycle (Hehenkamp *et al*., 2006, La Marca *et al*., 2006) in contrast to the other parameters, which can ony be determined in the early follicular phase. This study supported this phenomenon, because we did not see a change in the level of AMH after an acute endogenous rise in FSH (CCCT) and an acute exogenous rise in FSH (EFORT).
In women, serum AMH expression can first be observed in granulosa cells of primary follicles, and expression is strongest in preantral and small antral follicles. AMH expression disappears in follicles of increasing size and is almost lost in follicles larger than 8 mm, where only very weak staining remains, restricted to the granulosacells of he cumulus (Weenen *et al*., 2004). This expression pattern is in agreement with the observation that AMH plays a role in initial recruitment of and in the selection of the dominant follicle (Visser *et al*., 2006). Our observations, that serum AMH levels in contrast to those of inhibin B and E2 do not substantially alter after injection of 300 IE FSH or 5 days of clomiphene citrate, support the notion that secretion of this substance is probably a measure of the ovarian follicle population and not of the cyclic gonadotropic hormonal status of the patient.
In conclusion AMH is comparable with other commonly used ovarian reserve test, but is probably most applicable in general practice, because it can be measured throughout the cycle, an advantage for both patients and clinicians. The predictive value of AMH for poor response is comparable with that of AFC, but unfortunately not for the prediction of hyper responders. The great advantage of AFC over any other test is its potential usefulness for its ability to concomitantly predict low and high responders.

Table 9 Poor versus normal responders. Characteristics of the groups (values are means ± SD).

	Poor reponders N = 29	Normal responders N = 81	p
Baseline characteristics			
Age (y)	35.3 ± 3.0	33.5 ± 4.0	0.029
Cycle day 3			
FSH (IU/l)	12.0 ± 11.5	6.6 ±1.8	< 0.001
E2 (pmol/l)	124.1 ± 54.1	138.4 ± 156.5	0.632
Inhibin B (ng/l)	76.0 ± 47.4	93.1 ± 43.0	0.077
AMH (µg/l)	1.48 ± 2.59	3.53 ± 2.46	<0.001
Endpoints			
Total number of follicles	4.6 ± 2.6	17.7 ± 9.7	< 0.001
Total number of oocytes	3.0 ± 1.6	14.9 ± 8.1	< 0.001
Ongoing pregnancy	2 (10%)	22 (24%)	0.001

References

Broekmans FJ, Kwee J, Hendriks D, Mol BW , Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod Update 2006;12:685-718.

Durlinger AL, Kramer P, Karels B, de Jong FH, Uilenbroek JThJ, Grootegoed JA, Themmen APN. Control of primordial follicle recruitment by anti-Mullerian hormone in the mouse ovary. Endocrinology 1999;140:5789-96.

Durlinger AL, Visser JA, Themmen AP. Regulation of ovarian function: the role of anti-Mullerian hormone. Reproduction 2002;124:601-9.

Fanchin R, Schonauer LM, Righini C, frydman N, Frydman R, Taieb J. Serum anti-Mullerian hormone dynamics during controlled ovarian hyperstimulation. Hum Reprod 2003;18:328-32.

Hehenkamp JK, Loomans CWN, Themmen APN, de Jong FH, te Velde ER, Broekmans FJM. Anti-Mullerian Hormone levels in the spontaneous menstrual ccle do not show substantial fluctuation. J Clin Endocrinol Metab 2006;10:4057-63.

Kevenaar ME, Meerasahib MF, Kramer P, van de Lang-Born BM, de jong FH, Groome NP, Themmen AP, Visser JA. Serum anti-mullerian hormone levels reflect the size of the primordial follicle pool in mice. Endocrinology 2006;147:3228-34.

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB, Schoemaker J. Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study. Hum Reprod 2003;18:1422-7.

Kwee J, Schats R, McDonnell J, Schoemaker J, Lambalk CB. The Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of low and hyperresponders to in vitro fertilization (IVF). Fertil Steril 2006;85:1714-22.

Kwee J, Elting M, Schats R, McDonnell J, Lambalk CB. Ovarian volume and antral follicle count for the prediction of low and hyper responders with in vitro fertilization. Submitted.

La Marca A, Stabile G, Artenisio AC, Volpe A. Serum anti-Mullerian hormone throughout the human menstrual cycle. Hum Reprod 2006:21, 3103-7.

Latin-American Puregon IVF Study Group. A double-blind clinical trial comparing a fixed daily dose of 150 and 250 IU of recombinant follicle-stimulating hormone in women undergoing in vitro fertilization. Fertil Steril 2001;76:950-6.

Out HJ, Braat DM, Lintsen BME, Gurgan T, Bukulmez O, Gokmen O, Keles G, Caballero P, Gonzalez JM, Fabregues F et al. Increasing the daily dose of recombinant follicle stimulating hormone (Puregon®) does not compensate for the age-related decline in retrievable oocytes after ovarian stimulation. Hum Reprod 2000;15: 29-35.

Out HJ, David I, Ron-El R, Friedler S, Shalev E, Geslevich J, Dor J, Shulman A, Ben Rafael Z, Fisch B et al. A randomized, double-blind clinical trial using fixed daily dose of 100 or 200 IU of recombinant FSH in ICSI cycles. Hum Reprod 2001;16:1104-9.

Seifer DB, Mac Laughlin DT, Christian BP, Feng B, Shelden RM. Early follicular serum mullerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. Fertil Steril 2002;77:468-71.

Van der Meer M, Hompes PGA, de Boer JAM, Schats R, Schoemaker J. Cohort size rather than follicle-stimulating hormone threshold level determines ovarian sensitivity in polycystic ovary syndrome. J Clin Endocrinol Metab 1998;83:423-26.

Van Rooij IA, Broekmans FJ, te Velde ER, Fauser BC, Bancsi LF, de Jong FH, Themmen AP. Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod 2002;17:3065-71.

Van Rooij IA, Tonkelaar I, Broekmans FJ, Looman CW, Scheffer GJ, de Jong FH, Themmen AP, te Velde ER. Anti-mullerian hormone is a promising predictor for the occurrence of the menopausal transition. Menopause 2004;11:601-6.

Van Rooij IA, Broekmans FJ, Scheffer GJ, Looman CW, Habbema JD, de Jong FH, Fauser BJ, Themmen AP, te Velde ER. Serum antimullerian hormone levels best reflect the reproductive decline with age in normal women with proven fertility: A longitudinal study. Fertile Steril 2005;83: 979-87.

Visser JA, deJong FH, Laven JSE, Themmen APN. Anti-Mullerian hormone: a new marker for ovarian function. Reproduction 2006;1:1-9.

Weenen C, Laven JS, Von bergh AR, Cranfield M, Groome NP, Visser JA, Kramer P, Fauser BC, Themmen AP. Anti-Mullerian hormone expression pattern in the human ovary: potential implications for initial and cyclic follicle recruitment. Mol Hum Reprod 2004;10:77-83.

Chapter 7

A systematic review of tests predicting ovarian reserve and IVF outcome

F.J. Broekmans[1], J. Kwee[2], D.J. Hendriks[1], B.W. Mol[3] and C.B. Lambalk[2,]

[1] Department of Reproductive Medicine, University Medical Centre Utrecht, Utrecht, The Netherlands
[2] Division of Reproductive Endocrinology and Fertility and the IVF Centre, Department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands
[3] Centre for Reproductive Medicine, Department of Obstetrics and Gynecology, Academic Medical Centre, Amsterdam, The Netherlands

Human Reproduction Update 2006;6:685-718

Abstract

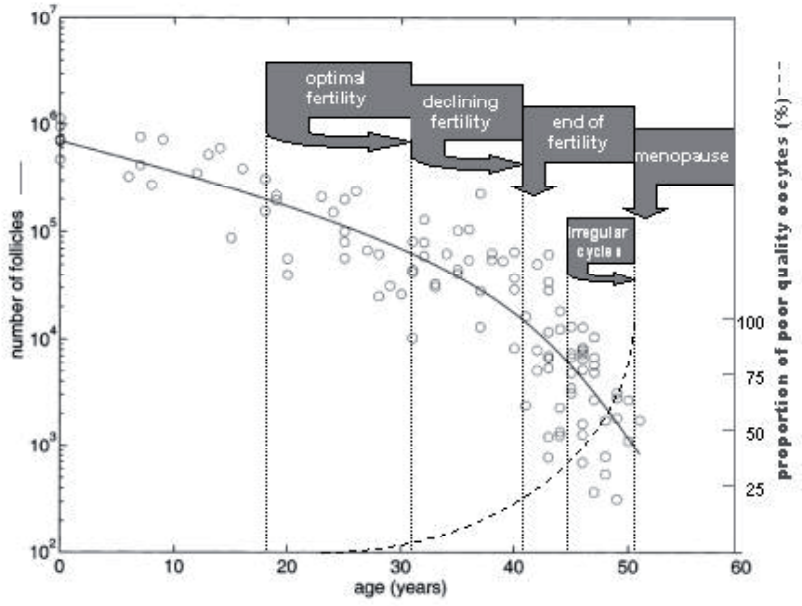
The age-related decline of the success in IVF is largely attributable to a progressive decline of ovarian oocyte quality and quantity. Over the past two decades, a number of so-called ovarian reserve tests (ORTs) have been designed to determine oocyte reserve and quality and have been evaluated for their ability to predict the outcome of IVF in terms of oocyte yield and occurrence of pregnancy. Many of these tests have become part of the routine diagnostic procedure for infertility patients who undergo assisted reproductive techniques. The unifying goals are traditionally to find out how a patient will respond to stimulation and what are their chances of pregnancy. Evidence-based medicine has progressively developed as the standard approach for many diagnostic procedures and treatment options in the field of reproductive medicine. We here provide the first comprehensive systematic literature review, including an a priori protocolized information retrieval on all currently available and applied tests, namely early-follicular-phase blood values of FSH, estradiol, inhibin B and anti-Müllerian hormone (AMH), the antral follicle count (AFC), the ovarian volume (OVVOL) and the ovarian blood flow, and furthermore the Clomiphene Citrate Challenge Test (CCCT), the exogenous FSH ORT (EFORT) and the gonadotrophin agonist stimulation test (GAST), all as measures to predict ovarian response and chance of pregnancy. We provide, where possible, an integrated receiver operating characteristic (ROC) analysis and curve of all individual evaluated published papers of each test, as well as a formal judgement upon the clinical value. Our analysis shows that the ORTs known to date have only modest-to-poor predictive properties and are therefore far from suitable for relevant clinical use. Accuracy of testing for the occurrence of poor ovarian response to hyperstimulation appears to be modest. Whether the a priori identification of actual poor responders in the first IVF cycle has any prognostic value for their chances of conception in the course of a series of IVF cycles remains to be established. The accuracy of predicting the occurrence of pregnancy is very limited. If a high threshold is used, to prevent couples from wrongly being refused IVF, a very small minority of IVF-indicated cases (~3%) are identified as having unfavourable prospects in an IVF treatment cycle. Although mostly inexpensive and not very demanding, the use of any ORT for outcome prediction cannot be supported. As poor ovarian response will provide some information on OR status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy.

Key words: IVF/ICSI outcome / ovarian reserve / ovarian stimulation

Introduction

In Western societies the introduction in the 1960s of reliable methods of contraception has led to the birth of fewer children per family. Driven by increasing levels of female education, a growing participation in labour force and career demands, postponement of childbearing has been a secondary consequence of the so-called sexual revolution (Leridon, 1998). These societal changes in family planning have caused a significant increase in the incidence of unwanted infertility due to female reproductive ageing (Weinstein *et al.*, 1993, Abma *et al.*, 1997, Ventura *et al.*, 2001).From studies on natural populations in which no consistent methods of birth control are applied, it has been shown that natural fertility starts to decline after the age of 30, accelerates in the mid-30s and will lead to sterility at a mean age of 41 (Spira, 1988, Wood, 1989, te Velde and Pearson, 2002) (Figure 1). The reduction in female fertility can also be shown from contemporary population studies. The chance of not conceiving a first child within one year increases from under 5% in women in their early 20s to approximately 30% or over in the age group of 35 years and older (Abma *et al.*, 1997). So, although the majority of women of older age will obtain the desired pregnancy within a one-year period, the chance of becoming subfertile increases ~6 fold in comparison with very young women.

Fig. 1. Quantitative (solid line) and qualitative (dotted line) decline of the ovarian follicle pool, which is assumed to dictate the onset of the important reproductive events. (Reproduced and adapted with permission from J.P. de Bruin and E.R. te Velde. Female reproductive aging: concepts and consequences. In Togas Tulandi and Roger G. Gosden, eds. Preservation of fertility. London, UK: Taylor&Francis, 2004: page 3)



The age-related effect on female fertility has also been shown in numerous reports on the results of IVF treatment in infertile couples. The probability of live birth obtained through IVF treatment clearly decreases after the age of 35 (Anonymous, 1995, Templeton *et al*., 1996) and the same has been shown to be true for the implantation rate per embryo (van Kooij *et al.*, 1996). In fact, female age has consistently been shown to be an important predictor of success in IVF treatment.

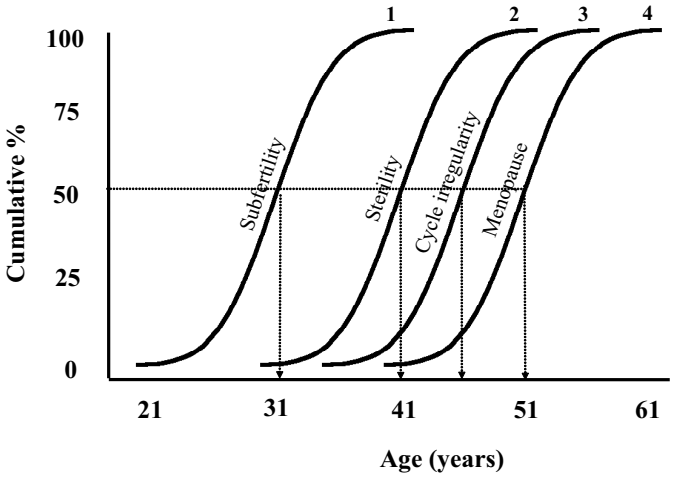
Over the past two decades, a number of so-called ovarian tests have been studied for their ability to predict outcome of IVF in terms of oocyte yield and occurrence of pregnancy. Some of these tests have become part of the routine diagnostic procedure for infertility patients that will undergo assisted reproductive techniques. With the current work we aim to provide an answer to the question of what the true value is of these tests to patient management. Evidence-based medicine has progressively developed as the standard approach for many diagnostic procedures and treatment options in the field of reproductive medicine (National Collaborating Center for Women's and Children's Health, 2004). Therefore, we provide a comprehensive systematic literature review, including an *a priori* protocolized information retrieval on all currently available and applied tests to determine ovarian reserve (OR).

What follows is first a general section in which we briefly outline the aims and the valuation of OR testing and the set-up of the systematic review. After this, we describe individually all currently available tests and their effectiveness with regard to prediction of ovarian response and pregnancy after IVF in generally accepted terms for diagnostic procedures. A unique feature of this systematic review is that we will furthermore provide where possible an integrated receiver operating characteristic (ROC) analysis and curve of all individual evaluated published papers of each test, as well as a formal judgement upon the clinical value.

The assessment of OR

OR can be considered normal in conditions where stimulation with the use of exogenous gonadotrophins will result in the development of at least 8–10 follicles and the retrieval of a corresponding number of healthy oocytes at follicle puncture (Fasouliotis *et al.*, 2000). With such a yield, the chances of producing a live birth through IVF are considered optimal. In general, as outlined earlier, age of the woman is a simple way of obtaining information on the extent of her OR, in terms of both quantity and quality (Templeton *et al.*, 1996). However, in the view of the substantial variation in the decline of reproductive capacity with age (te Velde and Pearson, 2002) (Figure 2), there is a need to identify women of relatively young age with clearly diminished reserve, as well as women around the mean age at which natural fertility on average is lost (41 years) but still with adequate OR. In clinical terms, we aim to identify women with a high risk of producing a poor response to ovarian stimulation and/or a very low probability of becoming pregnant through IVF, as well as those who still produce enough oocytes to have a good chance of becoming pregnant even if female age is advanced. If it appears possible to identify such categories of women, then management could be individualized, for instance by stimulation dose or treatment scheme adjustments (Tarlatzis *et al.*, 2003), by counselling against initiation of IVF treatment or pertinent refusal to accept initiation, or by indicating the necessity of early initiation of treatment before reserve has diminished too far.

Figure 2. Variations in age at the occurrence of specific stages of ovarian ageing. For explanation of the background of data, see te Velde and Pearson (te Velde, E. R. and Pearson, P. L. 2002). Reprinted with permission from te Velde and Pearson (te Velde, E. R. and Pearson, P. L. 2002)



OR is currently defined as the number and quality of the follicles left in the ovary at any given time. An accurate measure of the quantitative OR would involve the counting of all follicles present in both ovaries, as is done in post-mortem studies (Block, 1952). For obvious reasons, in OR testing, the true size of the follicle pool has not been used as the benchmark for evaluation (Lass *et al.*, 1997a, Lambalk *et al.*, 2004, Lass, 2004, Sharara and Scott, 2004), apart from one distinct study (Gulekli *et al.*, 1999), where whole ovary counts served as reference for several OR tests (ORTs). Instead, several proxy variables of the pool size are used in studies on diagnostic accuracy, like ovarian response to hyperstimulation with exogenous FSH in IVF and the occurrence of menopause or menopausal transition, as these events are quantitatively determined. Although related, the quality of the oocyte released from the dominant follicle at ovulation represents the other aspect of ovarian reserve. Proxy variables for oocyte quality currently used are the pregnancy probability in infertility treatment like IUI and IVF or in the follow-up of couples during and after the initial infertility work-up.

We should therefore realize that in the vast majority of studies on ORTs that will be discussed below, either ovarian response or occurrence of pregnancy in IVF serves as the benchmark to judge upon the accuracy and clinical value of the test under study. Ovarian response to adequate stimulation may be considered the most accurate, though still indirect, representation of the status of the primordial follicle pool, as it is a condition that is continuously present in the individual that undergoes the test. In contrast, the occurrence of pregnancy in such an individual may be influenced by many more factors than oocyte, and hence embryo quality, alone. Only if the occurrence of pregnancy is studied in a series of treatment cycles it may represent a solid proxy variable of the benchmark for ovarian reserve. Most ORTs are quite adequate in predicting ovarian response, but often fail to correctly predict the occurrence of pregnancy, especially if only one IVF cycle was studied.

Properties of test evaluation

ORT evaluation using response and/or pregnancy as reference or outcome variables should imply the assessment of predictive accuracy and clinical value of the test. Accuracy refers to the degree by which the outcome condition is predicted correctly. Summary statistics of accuracy include *sensitivity* (rate of correct identification of cases with poor response), *specificity* (rate of correct identification of cases without poor response), *likelihood ratio* (LR, how many times more likely particular test results are in patients with poor response than in those without poor response) and *diagnostic odds ratios* (DOR, the odds of positive test results in cases with poor response over the odds of positive test results in those without poor response) (Deeks, 2001, Grimes and Schulz, 2005). To identify all cases that will respond poorly to stimulation without judging many normal responders badly, the test must have high sensitivity and high specificity.

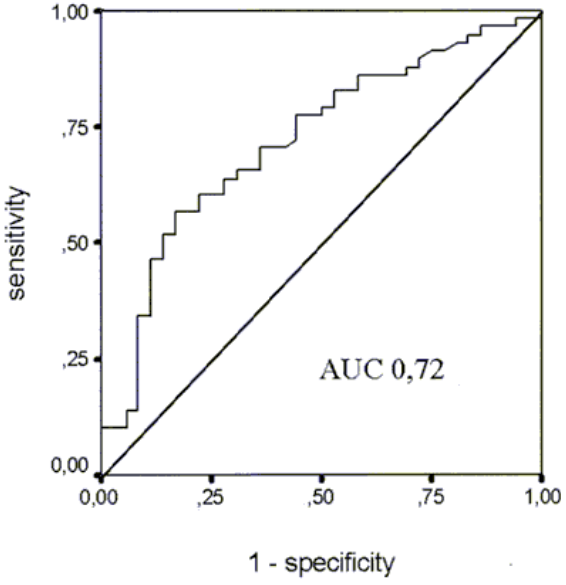
Positive LRs above 10 and negative LRs below 0.1 are considered as indicators of an adequate diagnostic test, while values between 5 and 10 and below 0.2 are considered to indicate a moderate test. As such, the LR can be considered a clinically useful tool to help judge the performance of the test, as the value will change when the threshold for an abnormal test is shifted.

The diagnostic odds ratio is an adequate measure when combining studies in a systematic review, as a single diagnostic odds ratio corresponds to a set of sensitivities and specificities depicted by an ROC curve and is considered threshold independent (Figure 3). It therefore can be considered a good parameter to compare the overall accuracy of a test evaluated in different studies. Although the DOR values will be higher for tests with better combinations for sensitivity and specificity, this value has not been advocated as a single measure of clinical value, as changes in the threshold used will not be expressed by a change in DOR value. For the meta-analytic approach, the range of DOR values across studies gives some indication as to the homogeneity of such studies.

Finally, the *area under the ROC curve* provides information on the overall discriminatory capacity of the test. Values of 1.0 imply perfect and that of 0.5 indicate completely absent discrimination.

Clinical value incorporates the question whether application of the test at a certain threshold will really change management or costs or safety or success rates on a population basis. It deals with the valuation of false positive and false negative test results in relation to the consequences of these test results for clinical decisions. Also it implies the rate of abnormal test results leading to altered decisions within the population of interest.

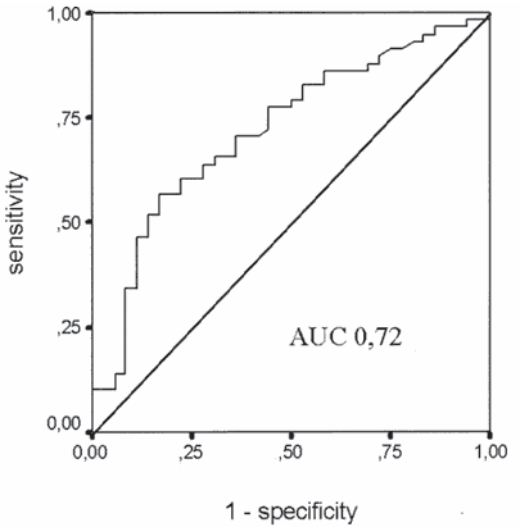
Figure 3. Receiver Operator Characteristics (ROC) curve depicting the continuous relationship between sensitivity and specificity with shifting cut off values for a given test. The area under the ROC curve (AUC) provides general information on the discriminatory capacity of the test.



Design of ORT studies

Studies on the predictive accuracy and clinical value of ORTs should preferably be prospective in design, should examine cohorts of patients in IVF settings without exclusion of cases with signs of diminished ovarian reserve and patient management should not have been influenced by the test under study (verification bias). Also, evaluation should be equally weighted for every case, thus every case should contribute the same amount of cycles to the analysis. In most studies, only one IVF cycle is studied. A case–control design for the purpose of OR testing bears the disadvantage of retrospection and the absence of a reliable estimate of disease prevalence. The tests under study should in principle be reproducible, both at the laboratory (hormone assays) and at the operator level (ultrasound examination). Also, the outcome of treatment (response and pregnancy), serving as the reference for ovarian reserve, should be clearly defined. The accuracy in predicting a certain outcome by the test under study should be evaluated by constructing contingency tables at several threshold levels for an abnormal test. Using the calculated sensitivity and specificity from each threshold level, a ROC curve (Figure 3) can be drawn and the calculated area under this curve represents the overall predictive accuracy of the test. Assessment of the clinical value is a complex process in which the applicability in daily practice should become clear. The overall accuracy represented by the ROC curve, the choice of a threshold for abnormality, the rate of abnormal tests at that threshold, the post-test probability of disease (i.e. poor response or non-pregnancy), the valuation of false positive and false negative test results and the consequence for patient management of an abnormal test will all contribute to the process of deciding whether a test is useful or not. Finally, the cost of carrying out the test as a routine measure and the burden to the patient balanced against the reduction in costs by excluding cases with low pregnancy prospects should contribute to the decision whether or not to apply a test.

Figure 3. Receiver Operator Characteristics (ROC) curve depicting the continuous relationship between sensitivity and specifity with shifting cut off values for a given test. The area under the ROC curve (AUC) provides general introduction on the discriminatory capacity of the test.



ORTs in relation to other predictors of success

It is important for patients who are considering treatment with IVF to know the probability of success in the course of a series of IVF treatment cycles. The possibility of a live birth for any couple undergoing treatment will depend on the success rate at the individual clinic. However, equally important in the prediction of outcome are the characteristics of the couple seeking treatment (Stolwijk *et al.*, 1996, Templeton *et al.*, 1996, Sharma *et al.*, 2002). Serious effort has been put into the build-up of prediction models that estimate the probabilities for success prior and during subsequent IVF cycles. In general, these models appeared inaccurate when external validation studies were carried out (Stolwijk *et al.*, 1998, Smeenk *et al.*, 2000). Intuitively, many IVF centres will use factors like female age, parity, duration of infertility, ovarian response in the first IVF attempt and embryo quality for individual counselling, albeit not through a formal prediction model. Within this practice, ORTs also may play a certain role and female age will be the one ORT applied almost without exception. The pressing question would be to what extent other, endocrine- or ultrasound-based, ORTs contribute and add to the prognostic information already obtained from the infertility work-up or the first IVF cycle. To date, studies specifically addressing this question are scarce or do not include the full range of prognostic factors available.

There are a number of studies (Eimers *et al.*, 1994, Collins *et al.*, 1995, Snick *et al.*, 1997, Hunault *et al.*, 2004, Hunault *et al.*, 2005) that offer a model, based on factors like duration of subfertility, female age, parity, sperm quality and post-coital test, for the prediction of live birth among untreated subfertile couples. However, none of these models included ORTs, apart from female age. Only one study showed that on top of predictions based on the Eimers model, ORTs failed to add relevant information to the couple's chances for a spontaneous pregnancy (van Rooij *et al.*, 2005).

General remarks on physiological background of ORTs

Tests that are used to predict some defined outcome related to ovarian reserve almost without exception give assessment of the number of follicles remaining at some time point in both ovaries. Any marker giving an estimate of the remaining pool will at the same time be capable of providing, to some extent, information on oocyte quality. But on average, from prediction studies it seems that some markers give a better indication of quality than others. Female age, for instance, is the basic factor that is related to both quantity and quality. Basal FSH, through the feedback of inhibin B and estradiol, will represent cohort size but mostly at the extremes and therefore give a more thorough indication of quality aspects. This is in contrast to the more direct quantitative tests using antral follicle count (AFC), anti-Müllerian hormone (AMH) and ovarian volume (OVVOL) that are capable of describing a more complete range of ovarian reserve states. By choosing the right thresholds these tests may eventually correctly predict oocyte quality. The true relation between quantity and quality, however, remains a source of debate. Quantity is an aspect of ovarian reserve that is present in a continuous state and therefore offers a more or less continuous measurability. Quality, however, comes to expression every now and then, even in the setting of IVF. The relationship between the two aspects of ovarian reserve has become more evident when the predictive value of a poor response in a first IVF cycle was examined towards the probability of pregnancy in the actual or subsequent cycles (Klinkert *et al.*, 2004). While cases with a normal response in additional cycles yielded acceptable rates of pregnancy, it was shown that in repeated poor responders this probability never surpassed 10% (de Boer *et al.*, 2002, Lawson *et al.*, 2003, Klinkert *et al.*, 2004). It is also important to remember that there are several factors that contribute to the occurrence of pregnancy other than ovarian reserve, such as embryo transfer technique and number of embryos replaced. Even in young women with normal reserve the chance of non-pregnancy remains at least at the 50% level. So, a non-pregnancy state after IVF may even be attributed to unknown, yet non-ovarian reserve related, factors.

Approach of the systematic review

The aim of the systematic review on the value of diagnostic tests is to obtain an overall estimate of the test accuracy and clinical value based on all present evidence, after assessing the quality of the included studies and evaluating the variation in findings among the studies (Irwig *et al.*, 1995, Deeks, 2001, Deville *et al.*, 2002, Honest and Khan, 2002, Glas *et al.*, 2003). Systematic review and meta-analysis on diagnostic accuracy and value implies consecutive steps as summarized in Table 1(Irwig *et al.*, 1994, Mol *et al.*, 1997) please see addendum.

For each study finally included in the meta-analysis, sensitivity and specificity are calculated from the contingency tables. Homogeneity of the sensitivity–specificity points is tested by means of the χ^2-test statistic. A summary point estimate of sensitivity and specificity and the 95% confidence interval is calculated if homogeneity cannot be rejected. In case of heterogeneity, logistic regression is used to evaluate whether Quality/Methodology characteristics of a study are associated with the discriminative capacity of the test under study.

Table 1. Stepwise approach to the systematic review and meta-analysis of diagnostic tests

1	Define the objective	Test and disease of interest. Reference standard for the disease. Impact of test result on clinical management. Comparison of tests.
2	Literature search	Search, link and MESH terms. In- and exclusion criteria. Databases used. Cross references. Contact authors for raw data if appropriate.
3	Data Extraction	Contingency table. Quality/Methodology characteristics. Extraction by two independent researchers. Disagreement solved by third independent researcher.
4	Heterogeneity Test	Chi square on Sensitivity (Sens) and Specificity (Spec) and provide ROC plot and Sens, Spec and Diagnostic Odds ratio (DOR) plot with 95% CI. Focus on outliers
5	Heterogeneity not rejected	Calculate Summary point estimates for Sens and Spec and 95% CI
	Heterogeneity rejected	Logistic regression analysis on relation Quality/Methodology characteristics and test accuracy. If present: subgroup analysis. If absent assume cut-off point effect.
6	Data Pooling	Spearman correlation between Sens and Spec (r< -0,5) or fixed effect logistic regression of lnDOR with an interaction term for test and study.
	Sens and Spec related and/or DORs homogenous	Summary ROC curve estimation using random-effects regression model.
	Sens and Spec not related and/or DORs heterogenous	No Pooling possible. Subgroup analysis?
7	Assess Clinical Value	Positive predictive value of abnormal test at various prevalence values using various cut offs based on Summary ROC curve, in correspondence with abnormal test rate. If no estimated curve or point: comparison of individual Sens and Spec points with desired level of Sens and Spec

If one of the study characteristics is found to have a statistically significant impact on the performance of the test, further analysis is performed in subgroups of patients. If not, it is explored whether the differences in sensitivity–specificity combinations are because of the use of different threshold levels of the test under study. For this purpose, a Spearman correlation coefficient is calculated to assess the association between sensitivity and specificity. If there is a negative correlation as defined by a correlation coefficient of –0.5 or stronger, the individual pairs of sensitivity and specificity are considered to originate from a single ROC curve. All sensitivity–specificity points are then plotted and a summary ROC curve is estimated using a random-effects regression model (Littenberg and Moses, 1993, Midgette *et al.*, 1993, Moses *et al.*, 1993).

An important issue is the fact that individual studies may produce highly variable sensitivity–specificity points in the ROC space. This is generally explained by variation in the applied threshold level for an abnormal test across the studies or the presence of considerable study heterogeneity. As in the formal analysis, the presence of heterogeneity in design will be dealt with, and the variation in sens/spec points is generally attributed to the variation in threshold levels and thus allows us to construct a summary ROC curve. At the same time, the threshold variation will prevent the possibility of assessing a single threshold for a specific test that has a generalizable value. This will only become possible if from every study the original database would be available and to date this seems to be an extreme effort.

To assess the clinical value of the test under study for the assessment of disease state (i.e. poor response or non-pregnancy), the positive and negative predictive values are calculated using the estimated summary ROC curve and assuming arbitrary prevalences of the disease in the population. An LR for a positive (or abnormal) test result is then calculated for each point on the estimated ROC curve. Subsequently, the post-test probabilities of disease at various LR values are then calculated for the arbitrary pre-test probabilities of disease, assuming independence between the pre-test probability and the performance of the test (Bancsi *et al.*, 2003). Final judgement depends on the overall accuracy, the choice of the test threshold, the post-test prediction at that threshold level and the valuation of a false positive test result. In case no estimated curve from the selected studies can be constructed, the judgement upon the clinical value is based on a comparison of a preset level of sensitivity and specificity with the observed levels in the various studies.

Systematic reviewing of ORTs

The aim of the present series of systematic reviews is to assess the true diagnostic accuracy and clinical value of the ORTs known to date, when applied in an IVF/ICSI population. Reference standards used to valuate the test properties are response to ovarian stimulation and occurrence of pregnancy. No preset definition was used for these standards. For every ORT under study, a computerized MEDLINE search was performed to identify articles on the subject outlined in the previous chapters published until December 2004. Checking of reference lists of articles already obtained was done, all in an iterative fashion. Keywords used for the various searches were ‘in vitro fertilization’ or ‘in vitro fertilisation’ or ‘assisted’ or ‘intracytoplasmatic’ or ‘intracytoplasmic’, in combination with ‘test-specific’ keywords, as mentioned in the tables.

One investigator *(DH or JK)* read all abstracts of the articles that were identified by the search. Any article reporting on the association of the test with poor ovarian response and/or non-pregnancy after IVF or possibly containing information that was to be transformed into a predictive tabulation was pre-selected. Subsequently, all pre-selected articles were fully read and judged independently by two investigators *(DH and JK)*, and separate 2 x 2 tables were constructed for cross classification of the test result and the occurrence of poor response and/or non-pregnancy, whenever possible. In the event of disagreement on the inclusion or exclusion of pre-selected studies for the meta-analysis or on the calculation of the 2 x 2 table data or the scoring of quality characteristics, the judgement of a third author *(FB or CL)* was decisive. Studies in which it was not possible to construct 2 x 2 tables were excluded. Cross-references in all selected articles were checked, and, if applicable, studies were added to the analysis. Each study was scored by the investigators on the following Quality/Methodology characteristics: (i) sampling (consecutive versus other), (ii) data collection (prospective versus retrospective), (iii) study design (cohort study versus case–control study), (iv) blinding (present or absent), (v) selection bias, (vi) verification bias, (vii) analysis on one or multiple cycles per couple and (viii) definition of outcome, poor response and pregnancy. In the following sections, the results of search, data extraction, quality and methodology assessment and meta-analysis of extracted data as outlined above are discussed for every ORT comprised in this review.

Basal FSH

Systematic review

Through the search and selection strategy, a total of 37 studies reporting on the capacity of basal FSH to predict poor ovarian response and/or non-pregnancy after IVF and which were suitable for data extraction and meta-analysis were identified (Scott *et al.*, 1989, Padilla *et al.*, 1990, Toner *et al.*, 1991, Khalifa *et al.*, 1992, Chan *et al.*, 1993, Ebrahim *et al.*, 1993, Fanchin *et al.*, 1994, Huyser *et al.*, 1995, Licciardi *et al.*, 1995, Smotrich *et al.*, 1995, Balasch *et al.*, 1996, Csemiczky *et al.*, 1996, Martin *et al.*, 1996, Pruksananonda *et al.*, 1996, Gurgan *et al.*, 1997, Chang *et al.*, 1998a, Evers *et al.*, 1998, Ranieri *et al.*, 1998, Sharif *et al.*, 1998, Bassil *et al.*, 1999, Hall *et al.*, 1999, Bancsi *et al.*, 2000, Chae *et al.*, 2000, Creus *et al.*, 2000, Fabregues *et al.*, 2000, Jinno *et al.*, 2000, Penarrubia *et al.*, 2000, Mikkelsen *et al.*, 2001, Nahum *et al.*, 2001, van der Stege and van der Linden, 2001, Esposito *et al.*, 2002, Chuang *et al.*, 2003, Fiçicioğlu *et al.*, 2003, Kwee *et al.*, 2003, Yanushpolsky *et al.*, 2003, Akande *et al.*, 2004, Erdem *et al.*, 2004). Characteristics of the included studies are listed in Table 2 As shown, there was a large diversity with regard to the various aspects of methodology and quality, and the definition of poor ovarian response. Logistic regression analysis indicated no significant association between any of these study characteristics and the predictive performance of basal FSH. For example, whether the design of the study was retrospective or prospective did not influence the prognostic capacity of basal FSH.

Accuracy of poor response prediction

The sensitivities and specificities, as well as the positive LRs of an abnormal test and the DORs for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 3, please see addendum. Sensitivity and specificity points, as plotted in Figure 4, were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for

specificity 0.001). Therefore, calculation of one summary point estimate for sensitivity and specificity was not meaningful for overall judgement of accuracy. The Spearman correlation coefficient for sensitivity and specificity was –0.87, which was judged to be sufficient to estimate a summary ROC curve (Figure 4).

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table 4, please see addendum. Again, sensitivity and specificity points plotted in Figure 5 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation coefficient for sensitivity and specificity was –0.82 and as such was sufficient to estimate a summary ROC curve (Figure 5).

Clinical value

Based on the summary ROC curves depicted in Figure 4, a range of positive LRs was calculated and for each ratio the pre-FSH test probability of poor response and non-pregnancy was converted into a post-FSH-test probability. Table 5, (please see addendum) depicts the probability of obtaining a certain FSH test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a maximum positive LR of 8, the post-FSH-test probability of poor response will approximate 70% if the pre-FSH-test probability is assumed to be as high as 20%. As is apparent from this table, the probability of obtaining a test result (FSH level) with an LR of ~8 is quite small. Table 3 shows that in women with an increased FSH level the probability of poor response only increases substantially (3-fold or more) in studies applying a high threshold level for FSH, resulting in a very limited number of patients with an abnormal test result.

Even more so, for prediction of non-pregnancy, the extremely high FSH levels that are necessary to obtain the moderate positive LR of ~5, leading to a post-test pregnancy rate of less than 5% based on a pre-test rate of 20%, again occur only in a very limited number of patients (Table 5). Beyond the coordinate defined by specificity 0.90 and sensitivity 0.20, the summary ROC curve almost runs parallel to the line of equality. This indicates that this segment of the curve is 100% uninformative (LR ~1).

All this leads to the conclusion that with the use of basal FSH in regularly cycling women, accuracy in the prediction of poor response and non-pregnancy is adequate only at very high threshold levels, but because of the very low numbers of abnormal tests has hardly any clinical value. Considering this along with a false positive rate of ~5%, the test will not be suitable as a diagnostic test to exclude patients, but only as screening test for counselling purposes and further diagnostic steps, in which a first IVF attempt may be the step of choice (Roberts *et al*, 2005).

Table 2. Characteristics of included studies on Basal FSH (computerised search using the test specific keyword follicle stimulating hormone and FSH)

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	FSH-assay
Scott et al.	Yes	No	Cycle	Not applicable	Clinical/ongoing	RIA: Leeco Diagnostics
Padilla et al.	No	No	Cycle	Not applicable	Clinical	RIA: Amersham Corp.
Toner et al.	No	No	Cycle/retrieval	< 2 foll. 16 mm.	Ongoing	RIA: Leeco Diagnostics
Khalifa et al.	No	No	Retrieval	Not applicable	Ongoing	RIA: Leeco Diagnostics
Ebrahim et al.	Yes	Yes	Cycle	< 3 oocytes	Term	RIA: Serono Diagnostics
Chan et al.	No	Not stated	Cycle	< 3 foll. 15 mm.	Clinical/ongoing	RIA: Diag. Products Inc.
Fanchin et al.	Yes	Yes	Cycle	< 3 oocytes	Not applicable	Immunometric: Kodak Diag.
Huyser et al.	No	Yes	Cycle	Not applicable	Term	IFMA: Delfia
Licciardi et al.	No	Not stated	retrieval	Not applicable	Ongoing	RIA: Leeco Diagnostics
Smotrich et al.	No	No	Cycle	< 2 foll. 16 mm.	Clinical	RIA: Nichols Inst. Radio.
Martin et al.	Yes	No	Cycle	Not applicable	Clinical	ACS-180: Chemilum.
Pruksanonda et al.	No	Yes	Cycle	< 3 foll.	Clinical	Fuorescense immunoassay
Csemiczky et al.	No	No	Cycle	Not applicable	Clinical	RIA: Diag. Products Inc.
Balasch et al.	Yes	Yes	Cycle	< 2 foll. 17 mm. or < 5 foll. 14 mm.	Not applicable	RIA: Immunotech Int.
Gurgan et al.	No	Yes	Cycle	< 2 foll. 18 mm.	Clinical	RIA: J&J Clin. Diagnostics
Sharif et al.	Yes	Yes	Cycle	< 4 foll. 14 mm.	Clinical	ACS-180: Chemilum.
Chang et al.	Yes	No	Cycle	Not applicable	Clinical	Not stated
Evers et al.	Yes	Yes	Cycle	< 4 foll. 14 mm.	Clinical	RIA : Delfia
Ranieri et al.	No	yes	Cycle	< 5 foll. 15 mm.	Not applicable	Immunometric: Nichols Inst.

Table 2. continued

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	FSH-assay
Hall et al.	No	No	Patient	Not applicable	Clincical	RIA
Bassil et al.	No	No	Cycle	Not applicable	Clinical	Not stated
Jinno et al.	Yes	No	Cycle	Not applicable	Not stated	Enzyme immunoassay: Abbott
Bancsi et al.	No	Yes	Cycle	Not applicable	Ongoing	Immunoan./immunometric: Chiron
Chae et al.	Yes	Yes	Cycle	Not applicable	Clinical	IRMA: Jeil Japan
Penarrubia et al.	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoezymometric: Technicon
Creus et al.	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoezymometric: Technicon
Fabregues et al.	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	IRMA: Immunotech
Mikkelsen et al.	Yes	No	Retrieval	Not applicable	Clinical	Immuno I; Bayer
Van de Stege et al.	Yes	Yes	Cycle	< 3 foll. 18 mm.	Clinical	RIA: Elecsys
Nahum et al.	Yes	No	Cycle	< 3 foll. 18 mm.	Clinical	MEIA : Abbott
Esposito et al.	No	Yes	Cycle	Not applicable	Live birth	Immuno I; Bayer
Fiçicioğlu et al.	Yes	Yes	Cycle	< 2 foll. or < 5 oocytes	Not applicable	ELISA: Serotec Ltd, UK
Chuang et al.	No	Yes	Cycle	Not stated	Ongoing	Chemilum. Immunoassay: Immulite
Yanushpolsky et al.	Yes	No	Retrieval	Not applicable	Delivery	Techn. Imm. Syst.: Bayer
Erdem et al.	Yes	No	Cycle	< 5 oocytes or < 3 foll. 18 mm.	Not applicable	Immunometric: Immulite 2000
Akande et al.	Yes	yes	Cycle	< 3 foll. 18 mm.	Not applicable	Immunofluorimetric: DELFIA
Kwee et al.	Yes	Yes	Cycle	Poor response <6 oocytes	Not applicable	Immunomet.: Amerlite/Delfia

Table 3. Performance of basal FSH in the prediction of **poor response** in **IVF patients** and shift from pre-test to post-test probability of poor response for patients with an abnormal (= lower than the cut-off) FSH result.

Author	Cycles (n)	FSH cut-off value (IU/l)	Prediction of poor response				Pre-FSH probability (%)	Post-FSH probability (%)	Proportion of patients/cycles with abnormal FSH (%)
			Sens*	Spec†	LR+	DOR			
Toner et al.	1,478	10	0.72	0.40	1.2	1.6	7	10	61
		15	0.45	0.75	1.8	2.4	7	15	27
		20	0.31	0.90	3.1	3.9	7	19	12
		25	0.22	0.96	5.5	6.7	7	29	5
Ebrahim et al.	111	11.5	0.80	0.93	11.4	49.0	5	38	11
Chan et al.	144	4.5	0.94	0.33	1.4	8.2	13	17	71
		6	0.72	0.71	2.5	6.3	13	27	35
Fanchin et al.	52	11	0.86	0.45	1.6	4.9	27	37	63
Smotrich et al.	292	15	0.00	0.95	0	2.8	2	0	4
Pruksanonda et al.	36	4	1.00	0.26	1.4	0.7	3	4	75
		8	1.00	0.71	3.5	4.7	3	10	31
Balasch et al.	120	ns	0.50	0.81	2.6	4.3	33	56	29
Gurgan et al.	637	10	0.47	0.82	2.6	2.9	16	33	23
		13	0.37	0.92	4.6	4.2	16	47	12
		15	0.33	0.95	6.6	4.9	16	56	9
		20	0.11	0.99	11.0	4.4	16	66	3
Sharif et al.	344	5.4	0.91	0.12	1.0	1.3	9	9	89
		10.8	0.31	0.93	4.4	5.9	9	31	9
Evers et al.	231	17	0.26	0.97	8.7	10.5	20	69	8
Ranieri et al.	177	9.5	0.81	0.65	2.3	8.2	27	48	47

Table 3. Continued.

Author	Cycles (n)	FSH cut-off value (IU/l)	Prediction of poor response				Pre-FSH probability (%)	Post-FSH probability (%)	Proportion of patients/cycles with abnormal FSH (%)
			Sens*	Spec†	LR+	DOR			
Penarrubia et al.	80	?	0.83	0.73	3.1	4.5	25	52	41
Creus et al.	120	9.45	0.65	0.81	3.4	11.0	33	67	35
Fabregues et al.	80	?	0.28	0.91	3.1	3.8	35	62	16
Van der Stege et al.	87	10	0.60	0.85	4.1	8.8	6	20	17
Nahum et al.	272	10	0.22	0.93	3.2	3.8	14	33	9
Fiçicioğlu et al.	58	7	0.76	0.76	3.1	9.9	43	70	47
Chuang et al.	1,045	10	0.32	0.87	2.4	3.1	9	19	15
Erdem et al.	32	?	0.63	0.81	3.3	9.7	50	77	41
Akande et al.	536	6	0.88	0.50	1.7	6.9	6	10	53
		9	0.59	0.87	4.5	9.7	6	22	16
		12	0.47	0.96	11.3	20.3	6	42	7
Kwee et al.	110	4	1.00	0.05	1.1	1.9	26	27	96
		6	0.93	0.40	1.5	8.8	26	36	69
		8	0.72	0.78	3.3	9.2	26	54	35
		10	0.34	0.96	9.3	8.97	26	77	12
		12	0.24	1.00	21.4	28.5	26	89	8

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio, ns = not specified.

Table 4. Continued.

Author	Cycles (n)	FSH cut-off value (IU/l)	Prediction of non-pregnancy				Pre-FSH probability (%)	Post-FSH probability (%)	Proportion of patients/cycles with abnormal FSH (%)
			Sens*	Spec†	LR+	DOR			
Pruksanonda *et al.*	36	4	0.78	0.50	1.6	3.6	89	93	75
		8	0.34	1.00	2.1	2.7	89	94	31
Csemiczky *et al.*	53	7	0.26	1.00	6.8	8.6	58	90	15
Gurgan *et al.*	637	10	0.24	0.80	1.2	1.2	81	84	23
		13	0.14	0.95	2.8	3.1	81	92	12
		15	0.11	0.97	4.3	4.6	81	95	8
		20	0.03	1.00	4.4	4.5	81	95	3
Sharif *et al.*	344	10.8	0.12	0.97	4.0	4.6	70	90	9
Chang *et al.*	149	10	0.13	0.97	4.3	5.5	74	92	10
Evers *et al.*	231	17	0.09	1.00	3.2	3.4	86	95	8
Hall *et al.*	110	9.4	0.77	0.27	1.1	1.95	39	40	75
		11.2	0.60	0.57	1.4	2.0	39	47	50
		13.3	0.33	0.81	1.7	2.0	39	52	25
Bassil *et al.*	83	10	0.45	0.10	0.5	0.1	92	85	49
		15	0.32	0.50	0.6	0.5	92	88	34
		20	0.09	0.80	0.5	0.4	92	85	10
		25	0.04	0.90	0.4	0.4	92	83	5
		30	0.03	1.00	0.5	0.5	92	83	3
Jinno *et al.*	271	15	0.05	0.96	1.1	1.1	65	67	4
Bancsi *et al.*	435	15	0.06	1.00	3.9	4.0	86	96	5
Chae *et al.*	118	8.5	0.46	0.85	3.0	4.6	89	96	42
Mikkelsen *et al.*	130	15	0.34	0.73	1.3	1.4	88	91	33
Van der Stege *et al.*	87	10	0.18	0.85	1.2	1.2	70	73	17
Nahum *et al.*	272	10	0.11	0.96	2.7	2.9	65	83	9
Esposito *et al.*	293	10	0.19	0.91	2.1	2.3	74	85	16
		11.4	0.11	1.00	8.9	9.9	74	96	8
Chuang *et al.*	1,045	10	0.18	0.91	2.0	2.2	70	82	15
Yanushpolsky *et al.*	483	10	0.22	0.88	1.9	2.1	62	75	18

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 4. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of basal FSH in the prediction of poor response. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

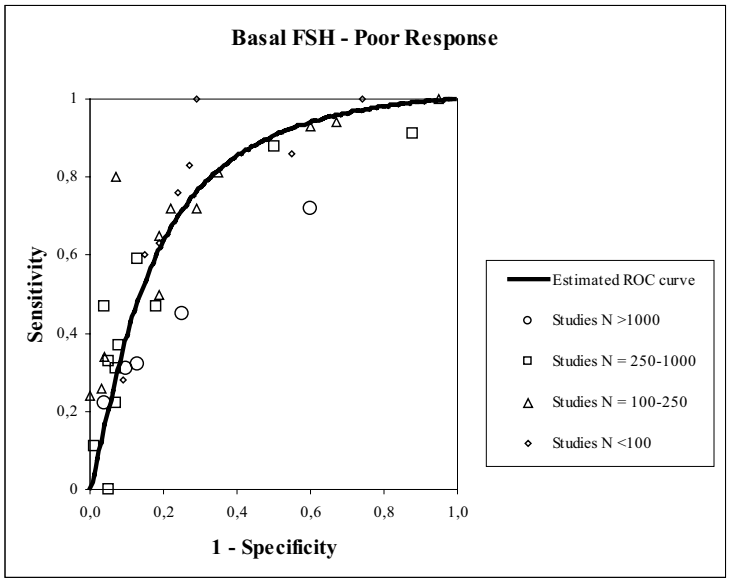


Figure 5. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of basal FSH in the prediction of non pregnancy. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

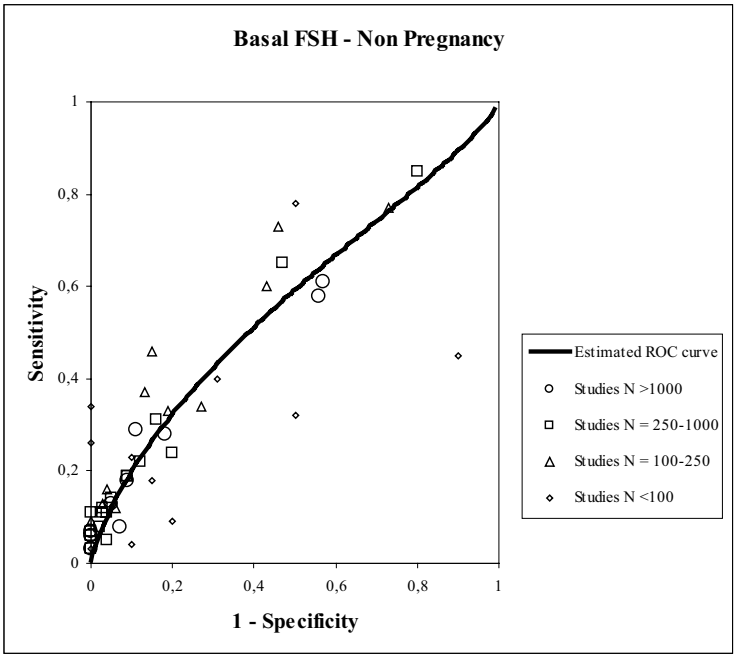


Table 5. The occurrence of the basal FSH results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results within this range (%)	Posttest probability poor response (%)	LR range	Occurrence of test results within in this range (%)	Posttest probability non-pregnancy (%)
0-1	68	<20	0-1	63	<80
1-2	15	20-33	1-2	22	80-89
2-3	8	33-43	2-3	9	89-93
3-4	3	43-50	3-4	1	93-94
4-5	2	50-56	4-5	1	94-95
5-6	1	56-60	5-6	1	95-96
6-7	1	60-64	6-7	1	96-96.5
7-8	1	64-67	7-8	1	96.5-97
>8	1	>67	>8	1	>97

AMH

Systematic review

Through the search and selection strategy, two studies reporting on the predictive capacity of AMH and which were suitable for data extraction and meta-analysis were identified (van Rooij *et al.*, 2002, Muttukrishna *et al.*, 2004). Characteristics of the included studies are listed in addendum, Table 6.

Accuracy of poor response prediction

The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 7, (see addendum) and in Figure 6. Homogeneity could not be rejected for sensitivity and specificity (χ^2-test statistic: *P*-value for sensitivity 0.12 and *P*-value for specificity 0.64), but this is merely because of the fact that only two studies were included. As can be seen from Figure 6, the points of the two studies can be thought of as originating from a single ROC curve (Spearman correlation coefficient between sensitivity and specificity is –0.81). The summary ROC curve that can be estimated from these points is also shown in Figure 6.

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy by AMH, as calculated from each study, are summarized in Table 8 As the study of Van Rooij was the only one detected, further meta-analysis is not useful. The ROC-curve derived from the data of Van Rooij *et al*. representing the accuracy of AMH in the prediction of non-pregnancy is shown in Figure 7.

Table 6. Characteristics of included studies on AMH (computerised search using the test specific keywords *anti-mullerian hormone* or *mullerian inhibiting factor* or *mullerian inhibiting substance*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/ Cancel	Definition Pregnancy	AMH-assay
Van Rooij *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll.	ongoing	Immuno-enzymometric (immunotech-Coulter)
Muttukrishna *et al.*	No	Yes	Cycle	< 4 foll. 15 mm.	not applicable	Immuno-enzymometric (immunotech)

Table 7. Performance of AMH in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal AMH result.

Author	Cycles (n)	AMH cut-off value (µg/l)	Prediction of poor response				Pre-AMH probability (%)	Post-AMH probability (%)	Proportion of patients/ cycles with abnormal AMH (%)
			Sens*	Spec†	LR+	DOR			
Van Rooij *et al.*	119	< 0.1	0.49	0.94	8.2	14,9	29	77	18
		< 0.2	0.54	0.90	5.7	11,3	29	70	23
		< 0.3	0.60	0.89	5.6	12,5	29	70	25
Muttukrishna *et al.*	69	< 0.1	0.76	0.88	6.6	24.9	25	68	28

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 8. Performance of AMH in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal AMH result.

Author	Cycles (n)	AMH cut-off value (µg/l)	Prediction of non-pregnancy				Pre-AMH probability (%)	Post-AMH probability (%)	Proportion of patients/cycles with abnormal AMH (%)
			Sens*	Spec†	LR+	DOR			
Van Rooij *et al.*	119	< 0.1	0.22	0.89	1.9	2.2	75	85	19
		< 0.2	0.27	0.85	1.8	2.1	75	84	24
		< 0.3	0.28	0.81	1.5	1.7	75	81	25

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 6. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **AMH** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

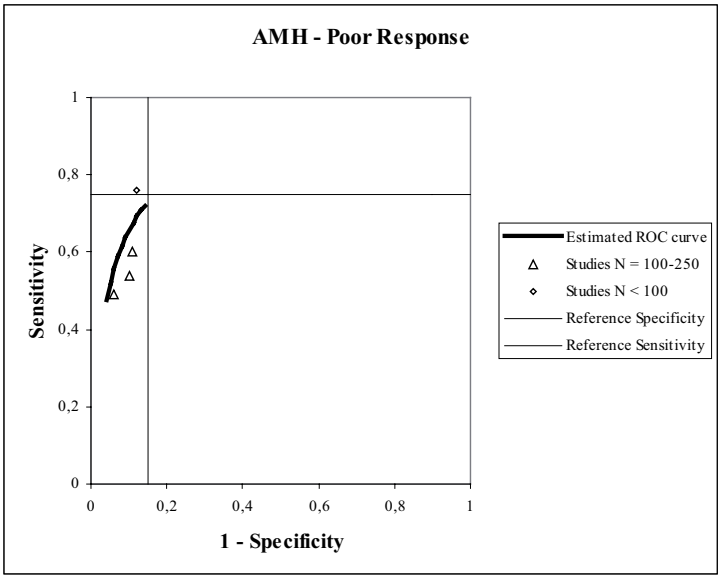
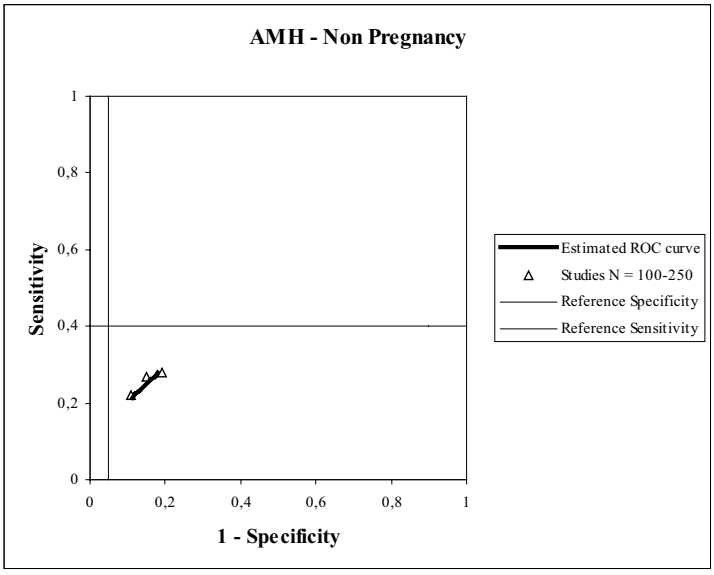


Figure 7. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **AMH** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points. Reference lines indicate a desired level of test performance.



Clinical value
As data from only two studies are available, it is not feasible to extract data on the interrelation between positive LRs, post-test probabilities and the rate of abnormal tests. However, looking at the performance of AMH in the prediction of poor response, a desired level for sensitivity of 75% and for specificity of 85% would imply that the test performs only moderately, especially at the sensitivity level. For non-pregnancy prediction, a desired level of sensitivity of 40% and specificity of 95% would imply that the test has hardly any value, unless very low threshold levels would be used, which will certainly lead to only very small percentages of abnormal tests. Additional studies are to be awaited to learn whether test capacity may prove to be more superior than current tests like basal FSH and the AFC (Hazout *et al.*, 2004, Muttukrishna *et al.*, 2005, Penarrubia *et al.*, 2005).

Inhibin B

Systematic review
We detected a total of nine studies reporting on the predictive capacity of inhibin-B and which were suitable for data extraction and meta-analysis (Balasch *et al.*, 1996, Seifer *et al.*, 1997, Hall *et al.*, 1999, Creus *et al.*, 2000, Fabregues *et al.*, 2000, Penarrubia *et al.*, 2000, Bancsi *et al.*, 2002a, Fiçicioğlu *et al.*, 2003, Erdem *et al.*, 2004). Characteristics of the included studies are listed in addendum Table 9 Variation among the definitions of poor response and study quality and design characteristics was clearly present but logistic regression analysis revealed that none of the items significantly impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Accuracy of poor response prediction
The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 10, see addendum. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics, as plotted in Figure 8, were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.002). The Spearman correlation coefficient for sensitivity and specificity was sufficient to estimate a summary ROC curve ($R = -0.93$, Figure 8). In the figure, it is clearly seen that all but one study were close to the estimated ROC curve, and that one study reported a clearly better accuracy (Fiçicioğlu *et al*, 2003). This study was of good quality, but reported on only a small number of patients.

Accuracy of non-pregnancy prediction
There were three studies that reported on the capacity of inhibin B to predict non-pregnancy. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table 11 Sensitivity and specificity as plotted in Figure 9 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.004 and *P*-value for specificity <0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.94, sufficient to estimate a summary ROC curve.

Table 9. Characteristics of included studies on inhibin B (computerised search using the test specific keyword *inhibin B*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Inhibin B assay
Balasch *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoenzymometric assay (Medgenix)
Seifer *et al.*	Yes	No	Patient	< 4 foll. 15 mm.	clinical	ELISA (Serotec Lim. UK)
Hall *et al.*	No	No	Cycle	Not applicable	clinical	ELISA (Serotec)
Creus *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Enzyme-linked immunosorbent (Serotec)
Fabregues *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoenzymatic (Medgenix)
Penarrubia *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoenzymometric (Immuno 1; Bayer)
Bancsi *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18 mm.	ongoing	Immuno-enzymometric (Serotec)
Fiçicioğlu *et al.*	No	Yes	Cycle	< 5 oocytes	Not applicable	ELISA (Serotec)
Erdem *et al.*	Yes	No	Cycle	< 5 oocytes (MII) or < 3 foll.	Not applicable	Immunosorbent (Serotec)

Table 10. Performance of inhibin B in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal inhibin B result.

Author	Cycles (n)	Inhibin B cut-off value (pg/ml)	Prediction of poor response				Pre-inhibin B probability (%)	Post-inhibin B probability (%)	Proportion of patients/ cycles with abnormal inhibin B (%)
			Sens*	Spec†	LR+	DOR			
Balasch *et al.*	120	ns	0.52	0.80	2.6	4.4	33	57	31
Seifer *et al.*	178	< 45	0.53	0.79	2.6	4.3	8	19	24
Creus *et al.*	120	ns	0.70	0.63	1.9	3.9	33	48	48
Fabregues *et al.*	80	ns	0.32	0.83	1.9	2.3	35	50	23
Penarrubia *et al.*	80	ns	0.89	0.29	1.3	3.6	25	30	76
Bancsi *et al.*	120	< 45	0.33	0.95	6.9	10	30	75	13
		< 53.8	0.39	0.94	6.5	10.1	30	74	16
Fiçicioğlu *et al.*	58	< 56	0.81	0.81	4.4	18.0	43	77	45
Erdem *et al.*	32	ns	0.69	063	1.8	3.7	50	65	53

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio, ns = not stated.

Table 11. Performance of the inhibin B in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal inhibin B result.

Author	Cycles (n)	Inhibin B cut-off value (pg/ml)	Prediction of non-pregnancy				Pre-inhibin B probability (%)	Post-inhibin B probability (%)	Proportion of patients/cycles with abnormal inhibin B (%)
			Sens*	Spec†	LR+	DOR			
Seifer *et al.*	178	< 45	0.28	0.92	3.5	4.5	79	93	24
Hall *et al.*	111	< 53.8	0.23	0.74	0.9	0.8	39	36	25
		< 76.5	0.60	0.56	1.4	1.9	39	46	50
		< 105.3	0.77	0.25	1.0	1.1	39	39	76
Bancsi *et al.*	120	< 45	0.17	1.00	5.2	6.1	78	94	13
		< 53.8	0.19	0.96	5.2	6.2	78	95	16

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 8. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **Inhibin B** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

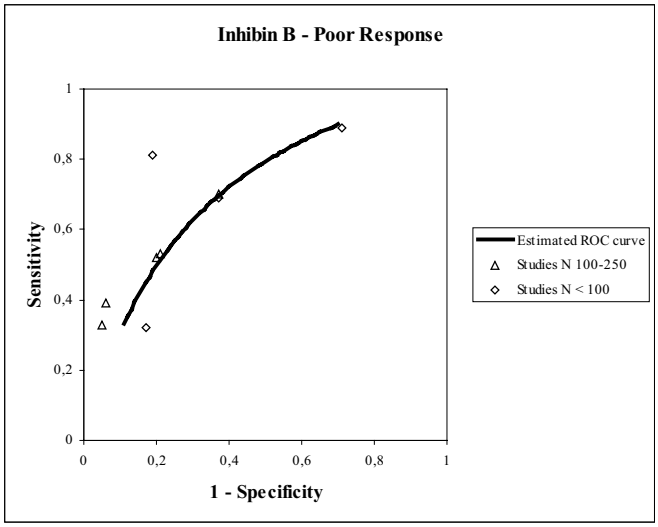
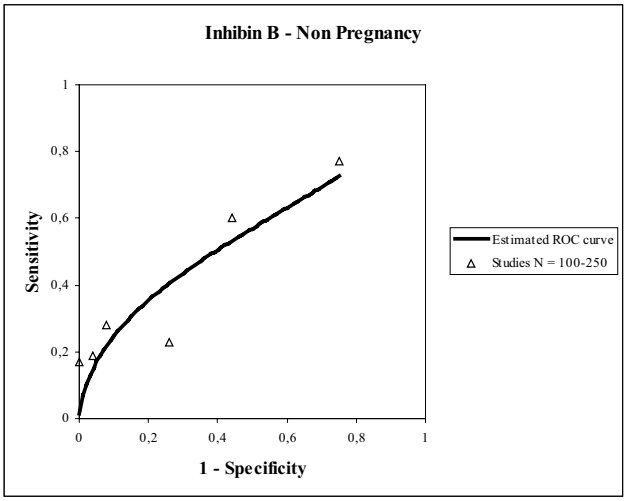


Figure 9. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **Inhibin B** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.



Clinical value

Based on the summary ROC curves depicted in Figure 8, a range of positive LRs was calculated and for each ratio pre-inhibin B-test probabilities of poor response or non-pregnancy (20 and 80%, respectively) were converted into post-inhibin B-test probabilities. Table 12 depicts the probability of obtaining a certain inhibin B test result and the corresponding LR, within different LR ranges for the prediction of poor response and non-pregnancy. At a very modest

LR of 4, the post-inhibin B-test probability of poor response will not be higher than 55%, while the chance of obtaining such a test result is very small.
For prediction of non-pregnancy, extreme threshold levels are necessary to obtain a modest positive likelihood ratio of ~4–5, leading to a post-test pregnancy rate of approximately 5%. Such abnormal test results occur only in a very limited number of patients, while the false positive rate will lead to unnecessary exclusions from IVF programs if the test is used in a diagnostic fashion.
With the use of basal inhibin B in regularly cycling women, the accuracy in the prediction of poor response and non-pregnancy is only modest at a very low threshold level. At best the test may be used as screening test for counselling purposes or to direct further diagnostic steps, like a first IVF attempt to observe the response to ovarian stimulation. Used in this way, the test may well be inferior to other tests discussed in this review.

Table 12. The occurrence of the inhibin B results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Posttest probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Posttest probability of non-pregnancy (%)
0-1	60	<20	0-1	79	<80
1-2	22	20-33	1-2	13	80-89
2-3	10	33-43	2-3	4	89-93
3-4	7,8	43-50	3-4	2	93-94
4-5	0,2	50-56	4-5	1	94-95
5-6	0	56-60	5-6	1	95-96
6-7	0	60-64	6-7	0	96-96.5
7-8	0	64-67	7-8	0	96.5-97
>8	0	>67	>8	0	>97

Basal estradiol

Systematic review
We detected a total of 10 studies reporting on the predictive capacity of basal estradiol and which were suitable for data extraction and meta-analysis (Licciardi *et al.*, 1995, Smotrich *et al.*, 1995, Evers *et al.*, 1998, Vazquez *et al.*, 1998, Hall *et al.*, 1999, Frattarelli *et al.*, 2000, Penarrubia *et al.*, 2000, Phophong *et al.*, 2000, Mikkelsen *et al.*, 2001, Ranieri *et al.*, 2001, Bancsi *et al.*, 2002a). Characteristics of the included studies are listed in addendum Table 13. Again, variation among the definitions of poor response and study quality and design characteristics was clearly present, but logistic regression analysis revealed that none of the items significantly impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Table 13. Characteristics of included studies on basal Estradiol (computerised search using the test specific keyword *estradiol*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Estradiol-assay
Smotrich *et al.*	No	No	Cycle	< 2 foll. 16 mm.	Clinical	RIA (Diag. Prod. USA)
Licciardi *et al.*	No	Not stated	Retrieval	Not applicable	Ongoing	RIA (Pantax South Monica, CA)
Ranieri *et al.*	No	Yes	Cycle	< 5 foll. 15 mm.	Not apllicable	RIA (Amersham Int. UK)
Evers *et al.*	Yes	Yes	Cycle	< 4 foll. 15 mm.	Clinical	RIA (Diag. Prod. USA)
Vazquez *et al.*					Clinical	
Hall *et al.*	No	No	Patient	Not applicable	Clinical	Enzyme immunoassay (Abott Lab. USA)
Frattarelli *et al.*	Yes	Yes	Cycle	< 3 foll.	Clinical	Immunolite immunoassay (Diag. Pord. USA)
Phophong *et al.*	Yes	Yes	Cycle	< 3 foll. 15 mm.	Clinical	RIA (Amersham Int. UK)
Penarrubia *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable	Immunoenzymometric (Immuno I; Bayer)
Mikkelsen *et al.*	Yes	No	Retrieval	Not applicable	Clinical	Autoanalyser (Immuno I; Bayer Denmark)
Bancsi *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18 mm.	Ongoing	AxSYM immunoanalyser (Abott Lab USA)

Table 14. Performance of basal estradiol in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal estradiol result.

Author	Cycles (n)	Estradiol cut-off value (pmol/l)	Prediction of poor response				Pre- Estradiol probability (%)	Post- Estradiol probability (%)	Proportion of patients/cycles with abnormal Estradiol (%)
			Sens*	Spec†	LR+	DOR			
Smotrich *et al*	292	> 294	0.83	0.92	10.8	60.0	2	19	9
		> 367	0.83	0.97	23.8	138.0	2	33	5
Ranieri *et al.*	177	> 350	0.79	0.81	4.1	15.8	27	60	36
Evers *et al.*	213	> 220	0.26	0.96	6.5	8.5	16	56	8
Vazquez *et al.*	248	> 92	0.64	0.38	1.0	1.1	9	9	62
		> 184	0.27	0.71	0.9	0.9	9	8	29
		> 275	0.09	0.88	0.7	0.7	9	7	12
		> 367	0.05	0.94	0.7	0.7	9	7	6
Frattarelli *et al.*	2,476	> 73	0.76	0.13	0.9	0.5	14	12	86
		> 147	0.34	0.56	0.8	0.7	14	11	43
		> 220	0.14	0.88	1.1	1.2	14	15	13
		> 294	0.06	0.97	1.95	2.0	14	24	4
		> 367	0.03	0.98	2.2	2.2	14	26	2
Phophong *et al.*	305	> 250	0.12	0.86	0.8	0.8	9	7	14
Penarrubia *et al.*	80	> ?	0.70	0.32	1.0	1.1	25	25	69
Bancsi *et al.*	120	> 200	0.31	0.74	1.2	1.2	30	33	28
		> 250	0.22	0.92	2.7	3.1	30	53	13

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 15. Performance of the basal estradiol in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal Estradiol result.

Author	Cycles (n)	Estradiol cut-off value (IU/l)	Prediction of non-pregnancy				Pre- Estradiol probability (%)	Post- Estradiol probability (%)	Proportion of patients/cycles with abnormal Estradiol (%)
			Sens*	Spec†	LR+	DOR			
Smotrich *et al.*	292	> 294	0.12	0.96	3.1	3.4	65	85	9
		> 367	0.08	1.00	8.7	9.4	65	94	5
Licciardi *et al.*	452	> 110	0.76	0.37	1.2	1.9	81	84	73
		> 165	0.42	0.69	1.3	1.6	81	85	40
		> 220	0.20	0.87	1.6	1.8	81	87	19
		> 275	0.08	1.00	7.4	8.0	81	97	7
Evers *et al.*	213	> 220	0.09	1.00	3.2	3.4	85	94	8
Vazquez *et al.*	248	> 92	0.60	0.33	0.9	0.8	70	67	62
		> 184	0.29	0.72	1.0	1.0	70	70	29
		> 275	0.11	0.85	0.7	0.7	70	63	12
		> 367	0.05	0.91	0.5	0.5	70	53	6
Hall *et al.*	120	> 108	0.71	0.25	0.95	0.8	38	36	73
		> 136	0.47	0.49	0.92	0.9	38	36	49
		> 167	0.20	0.72	0.7	0.6	38	30	25
Frattarelli *et al.*	2,476	> 73	0.84	0.12	0.96	0.8	54	53	86
		> 147	0.41	0.55	0.9	0.9	54	52	43
		> 220	0.12	0.87	1.0	0.99	54	54	13
		> 294	0.04	0.97	1.3	1.3	54	61	4
		> 367	0.02	0.99	1.9	1.95	54	70	2
Phopong *et al.*	305	> 250	0.13	0.83	0.8	0.8	77	72	14
Mikkelsen *et al.*	132	> 200	0.22	1.00	3.9	4.7	89	96	20
Bancsi *et al.*	120	> 200	0.27	0.70	0.9	0.9	78	76	28
		> 250	0.12	0.85	0.8	0.8	78	73	13

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Accuracy of poor response prediction

There were eight studies that reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 14. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics as plotted in Figure 10 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.002). The Spearman correlation coefficient for sensitivity and specificity was –0.50. As can be seen from Figure 10, this can be because of three outliers, which were extracted from the studies of Smotrich *et al*. and Ranieri *et al*. From neither the clinical nor the methodological point of view could a clear explanation be provided for the outliers. When correlation between sensitivity and specificity was assessed after exclusion of the three outliers, we found a very strong correlation (–0.94). Figure 10 shows two estimates of a summary ROC curve, one constructed with all data and one constructed after exclusion of the two studies with outlying data (Figure 10).

Figure 10. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **basal Estradiol** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

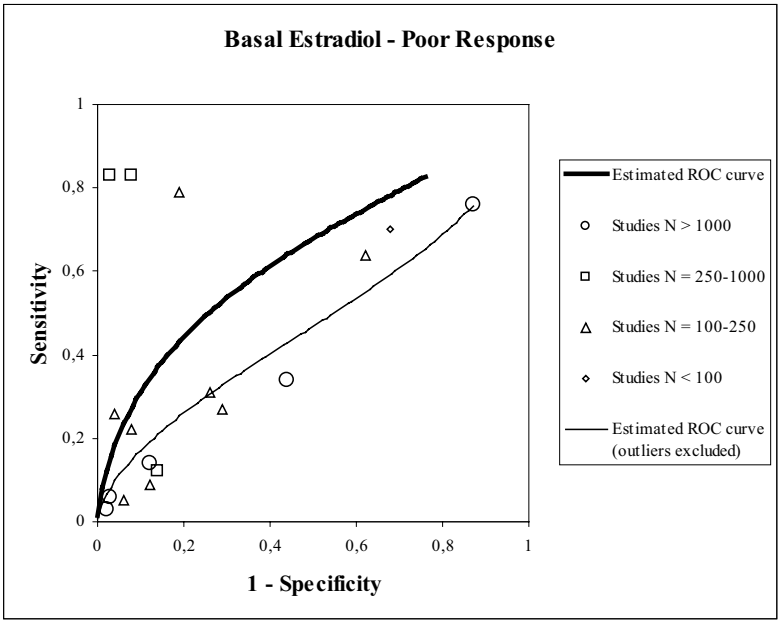
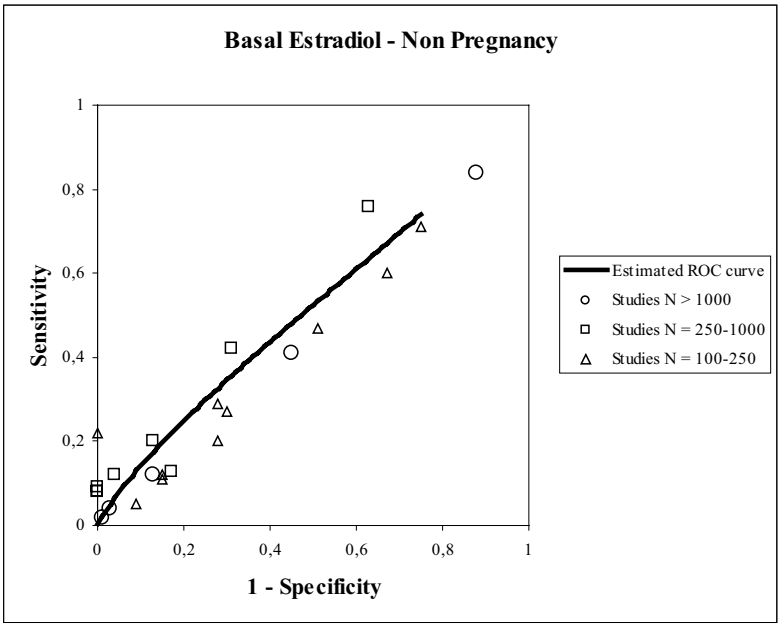


Figure 11. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **basal Estradiol** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.



Accuracy of non-pregnancy prediction

There were nine studies that reported on the capacity of basal estradiol to predict non-pregnancy after IVF. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table 15 Again, sensitivity and specificity as plotted in Figure 11 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity <0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.89, sufficient to estimate a summary ROC curve (Figure 11). This summary ROC curve is almost parallel to the line $x = y$, indicating virtually no discriminative capacity.

Clinical value

Based on the two summary ROC curves for all studies depicted in Figure 10, a range of positive LRs was calculated and for each ratio, pre-estradiol-test probabilities of poor response or non-pregnancy (20 and 80%, respectively) were converted into post-estradiol-test probabilities. Table 16 (please see addendum) depicts the probability of obtaining a certain estradiol-test result and the corresponding LR, within different LR ranges for the prediction of poor response and non-pregnancy. At a moderate LR of 4–5, the post-estradiol-test probability of poor response will not be higher than ~50%, while the chance of obtaining such a test result is very small.

For prediction of non-pregnancy no clear threshold levels can be identified for basal estradiol that will lead to an adequate combination of LR, post-test probability and abnormal test rate. This could be anticipated from the shape of the ROC curve in Figure 11

All this leads to the conclusion that the clinical applicability for basal estradiol as a test before starting IVF is prevented by the very low predictive accuracy, both for poor response and non-pregnancy.

AFC

Systematic review

Through the search and selection strategy, a total of 15 studies reporting on the predictive capacity of basal AFC and suitable for data extraction and meta-analysis were identified (Chang *et al.*, 1998b, Frattarelli *et al.*, 2000, Ng *et al.*, 2000, Sharara and McClamrock, 2000, Hsieh *et al.*, 2001, Nahum *et al.*, 2001, Bancsi *et al.*, 2002a, Erdem *et al.*, 2002, Fisch and Sher, 2002, Fiçicioğlu *et al.*, 2003, Frattarelli *et al.*, 2003, Jarvela *et al.*, 2003, Kupesic *et al.*, 2003, Yong *et al.*, 2003, Durmusoglu *et al.*, 2004). Characteristics of the included studies are listed in addendum Table17 Variation among the definitions of poor response and study quality and design characteristics is clearly present but logistic regression analysis revealed that none of the items significantly impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Table 16. The occurrence of the basal Estradiol results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			**Prediction of non-pregnancy (pre-test probability = 80%)**		
LR range	Occurrence of test results in this range (%)	Posttest probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Posttest probability of non-pregnancy (%)
0-1	83	<20	0-1	82	<80
1-2	12	20-33	1-2	17	80-89
2-3	3	33-43	2-3	1	89-93
3-4	1	43-50	3-4	0	93-94
4-5	1	50-56	4-5	0	94-95
5-6	0	56-60	5-6	0	95-96
6-7	0	60-64	6-7	0	96-96.5
7-8	0	64-67	7-8	0	96.5-97
>8	0	>67	>8	0	>97

Table 17. Characteristics of included studies on basal AFC (computerised search using test specific keywords *antral follicle count* or *antral follicle number*)

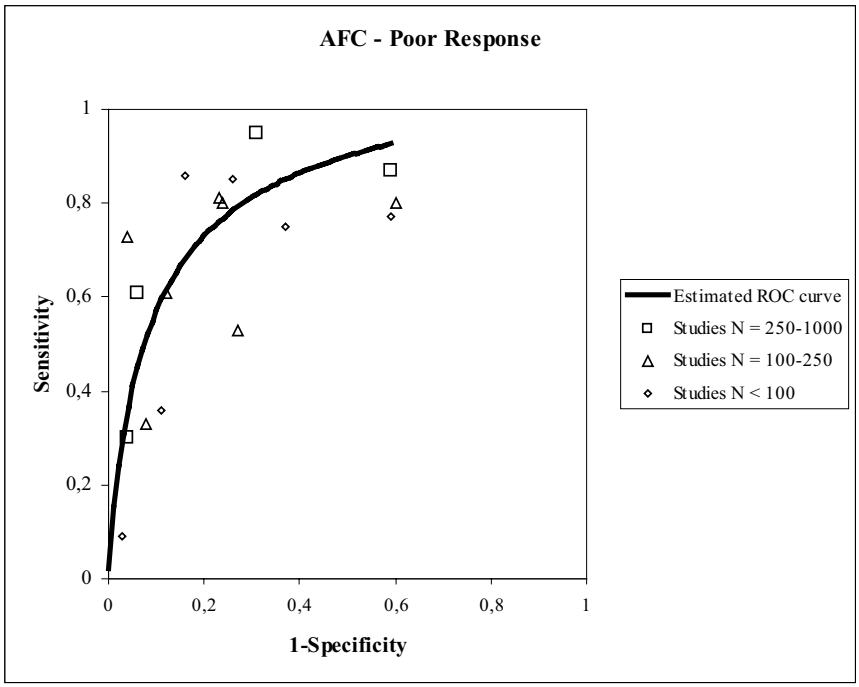
Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Diameter follicles (mm)	ultrasonograph
Chang et al.	Yes	No	Cycle	< 2 foll. 18 mm.	Ongoing	2-5	Accuson 120XP/10: 7 MHz probe
Ng et al.	Yes	Yes	Cycle	< 3 foll. 15 mm.	Clinical	Not stated	Aloka SSD-620: 5 MHz probe
Frattarelli *et al.* (2000)	No	Yes	Cycle	< 3 foll.	Not applicable	2-10	Acuson 128: 7 MHz probe
Sharara et al.	Yes	No	Cycle	Not stated	Clinical	2-8	not stated
Hsieh et al.	Yes	No	Cycle	No oocytes or poor foll. growth	Clinical	2-10	Acuson Aspen: 4 MHz probe
Nahum et al.	Yes	No	Cycle	< 3 foll. 18 mm.	Clinical	2-6	General electric RT-X200: 6.5 MHz probe
Fisch et al.	Yes	Yes	Cycle	Not applicable	Clinical	Not stated	not stated
Bancsi et al.	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18 mm.	clinical/ongoing	2-5	Toshiba Capasee SSA-220A: 7.5 MHz probe
Frattarelli et al. (2003)	Yes	Yes	Cycle	< 3 foll.	Not applicable	2-10	Acuson 128: 7 MHz probe
Järvelä et al.	Yes	Yes	Cycle	< 4 foll	Clinical	2-5	Kretz Combison 530D
Kupesic et al.	Yes	Yes	Cycle	Not applicable	Clinical	Not stated	Combison 530D: 7.5 MHz probe
Yong et al.	No	Yes	Cycle	< 4 oocytes or cancel	Clinical	2-10	Toshiba Eccocee: 7 MHz probe
Fiçicioğlu et al.							
Erdem et al.	Yes	Yes	Cycle	< 3 foll. 14 mm. or < 5 oocytes (MII)	Not applicablen	not stated	Aloka SSD-1000: 5 MHz probe
Durmusoglu et al.	No	No	Cycle	Poor foll. growth or < 3 oocytes (MII)	Not applicable	2-10	GE Logiq200: 6.5 MHz probe

Table 18. Performance of basal AFC in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal AFC result.

Author	Cycles (n)	AFC cut-off value (n)	Prediction of poor response Sens*	Spec†	LR+	DOR	Pre-AFC probability (%)	Post-AFC probability (%)	Proportion of patients/cycles with abnormal AFC (%)
Chang *et al.*	149	< 3	0.73	0.96	19.7	65	10	69	11
Ng *et al.*	128	< 4	0.33	0.92	4.2	5.7	2	9	9
		< 6	0.80	0.76	3.3	13	2	11	27
		< 9	0.80	0.40	1.3	2.7	2	5	61
Frattarelli *et al.* (2000)	278	< 10	0.87	0.41	1.5	4.7	8	12	61
Sharara *et al.*	127	< 4	0.53	0.73	1.9	3.0	15	26	31
Hsieh *et al.*	372	< 3	0.61	0.94	10.0	23	5	34	9
Nahum *et al.*	272	< 6	0.95	0.69	3.1	42	14	33	39
Bancsi *et al.*	120	< 4	0.61	0.88	5.1	12	30	69	27
		< 6	0.81	0.77	3.6	14	30	60	40
Frattarelli *et al.* (2003)	267	< 4	0.30	0.96	7.4	10	9	41	6
Järvelä *et al.*	45	< 4	0.86	0.84	5.4	32	16	50	27
Yong *et al.*	47	< 4	0.09	0.97	3.3	3.2	23	50	4
		< 6	0.36	0.89	3.3	4.6	23	50	17
Fiçicioğlu *et al.*	58	< 7	0.77	0.41	1.3	2.3	43	50	66
Erdem *et al.*	32	?	0.75	0.63	2.0	5.1	50	67	56
Durmusoglu *et al.*	91	< 6.5	0.85	0.74	3.3	16	26	53	41

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 12. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of the **AFC** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.



Accuracy of poor response prediction

The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 18. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics as plotted in Figure 12 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation coefficient for sensitivity and specificity was –0.57 and was judged to be sufficient to estimate a summary ROC curve (Figure 12).

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table 19. Again, sensitivity and specificity as plotted in Figure 13 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.66, sufficient to estimate a summary ROC curve (Figure 13).

Clinical value

Based on the summary ROC curves depicted in Figure 12, a range of positive LRs was calculated and for each ratio pre-AFC test probabilities of poor response or non-pregnancy were converted into a post-AFC-test probability. Table 20 depicts the probability of obtaining

a certain AFC test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a maximum positive LR of ~8, the post-AFC test probability of poor response will approximate 70%, if the pre-AFC-test probability is assumed to be as high as 20%. The probability of obtaining a test result (AFC) with a likelihood ratio ~8 is high enough to consider the AFC as a clinically valuable test for poor response prediction.
For prediction of non-pregnancy, the extremely low AFC that is necessary to obtain a moderate positive likelihood ratio of ~5, leading to a post-test pregnancy rate of less than 5% based on a pre-test rate of 20%, occurs only in an extremely limited number of patients (Table 20). Beyond the coordinate defined by specificity 0.80 and sensitivity 0.30, the summary ROC curve almost runs parallel to the line of equality. This indicates that this segment of the curve is 100% uninformative (LR ~1).
Based on these data, it can be concluded that the accuracy of the AFC for predicting poor response in regularly cycling women is adequate at a low threshold level, but because of the very limited numbers of abnormal tests has hardly any clinical value for pregnancy prediction. Added to the false positive rate of ~5% the test will not be suitable as diagnostic test to exclude patients on the basis of the presumed diagnosis of advanced ovarian ageing. It may well be used as a screening test for possible poor responders and for directing further diagnostic steps like a first IVF attempt, where the ovarian response to hyperstimulation will provide additional information (Hendriks *et al.*, 2005d).

Figure 13. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of the **AFC** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

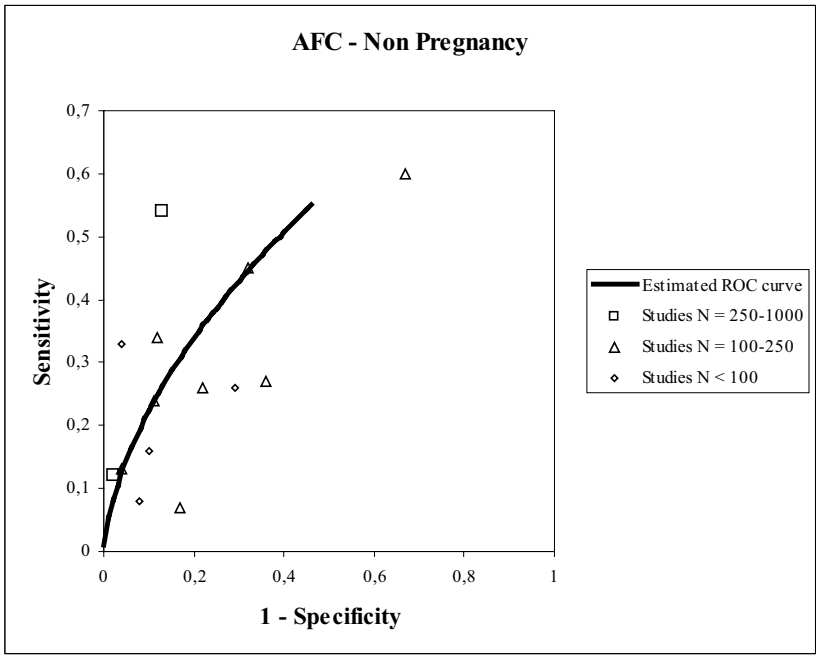


Table 19. Performance of basal AFC in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal AFC result.

Author	Cycles (n)	AFC cut-off value (n)	Prediction of non-pregnancy				Pre-AFC probability (%)	Post-AFC probability (%)	Proportion of patients/ cycles with abnormal AFC (%)
			Sens*	Spec†	LR+	DOR			
Chang *et al.*	149	< 3	0.13	0.96	3.6	3.6	83	94	11
Ng *et al.*	128	< 4	0.07	0.83	0.4	0.4	86	73	9
		< 6	0.26	0.78	1.2	1.2	86	88	26
		< 9	0.60	0.33	0.9	0.7	86	61	85
Sharara *et al.*	127	< 4	0.27	0.64	0.8	0.7	56	49	31
Hsieh *et al.*	372	< 3	0.12	0.98	6.9	6.7	68	94	9
Nahum *et al.*	272	< 6	0.54	0.87	4.0	7.9	64	88	39
Fisch *et al.*	200	< 10	0.24	0.89	2.2	2.6	59	76	19
Bancsi *et al.*	107	< 4	0.34	0.88	2.9	3.8	68	86	27
		< 6	0.45	0.68	1.4	1.7	68	75	41
Järvelä *et al.*	45	< 4	0.26	0.71	0.9	0.9	69	67	27
Kupesic *et al.*	56	< 4	0.33	0.96	8.3	11.8	61	92	22
Yong *et al.*	47	< 4	0.08	0.92	0.9	1.0	76	75	9
		< 6	0.16	0.90	1.6	1.7	79	86	27

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 20. The occurrence of the AFC results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			**Prediction of non-pregnancy (pre-test probability = 80%)**		
LR range	Occurrence of test results in this range (%)	Posttest probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Posttest probability of non-pregnancy (%)
0-1	68	<20	0-1	77	<80
1-2	10	20-33	1-2	16	80-89
2-3	4	33-43	2-3	5	89-93
3-4	6	43-50	3-4	0	93-94
4-5	0	50-56	4-5	2	94-95
5-6	0	56-60	5-6	0	95-96
6-7	0	60-64	6-7	0	96-96.5
7-8	0	64-67	7-8	0	96.5-97
>8	12	>67	>8	0	>97

OVVOL

Systematic review

For assessing the predictive value of OVVOL, the search detected a total of 10 studies available for data extraction and meta-analysis. Of these, two studies reported solely on the prediction of poor response (Sharara and McClamrock, 1999, Fiçiciolu *et al.*, 2003) and eight studies reported on the prediction of both poor response and pregnancy (Syrop *et al.*, 1995, Lass *et al.*, 1997b, Frattarelli *et al.*, 2000, Schild *et al.*, 2001, Bancsi *et al.*, 2002a, Jarvela *et al.*, 2003, Kupesic *et al.*, 2003, Erdem *et al.*, 2004). Study characteristics of the included studies are listed in addendum Table 21. Selection bias was present in almost half of all studies (Lass *et al.*, 1997b, Frattarelli *et al.*, 2000, Kupesic *et al.*, 2003, Erdem *et al.*, 2004). In three studies, patients were selected by basal FSH level (Frattarelli *et al.*, 2000, Kupesic *et al.*, 2003, Erdem *et al.*, 2004) and in the study by Lass *et al*. (Lass *et al.*, 1997b) only patients aged >36 years with an FSH level <15 IU/L were included. Three studies showed evidence of verification bias (Jarvela *et al.*, 2003, Kupesic *et al.*, 2003, Erdem *et al.*, 2004), implying that smaller OVVOL altered the management of the patient by applying higher FSH dosages.

Accuracy of poor response prediction

Sensitivities and specificities, positive LR and the DOR for the prediction of poor ovarian response are summarized in Table 22. Homogeneity for both sensitivity and specificity had to be rejected (χ^2-test: both *P*-values <0.001). Hence, the calculation of a summary point estimate for sensitivity and specificity was not meaningful. None of the study characteristics recorded had a statistically significant impact on the reported predictive performance of OVVOL. The Spearman correlation coefficient for the relation between sensitivity and specificity was –0.55, sufficient to estimate a summary ROC curve. This curve showed a modest overall predictive accuracy as can be seen in the ROC space in Figure 14.

Accuracy of non-pregnancy prediction

For the prediction of non-pregnancy, test characteristics for each study are summarized in Table 23. As with the data for ovarian response, homogeneity for sensitivity had to be rejected. However, specificity appeared to be homogeneous (χ^2-test: *P*-value 0.11). Because for the estimation of one summary point for sensitivity and specificity statistical homogeneity, both test parameters are required, this solution was abandoned. Logistic regression analysis showed that three studies which suffered from verification bias reported a significantly different accuracy compared to the seven remaining studies (p-value: 0.01). None of the other study characteristics had a significant impact on the estimates of test accuracy. In the subgroup analysis of the seven studies without verification bias, homogeneity was again rejected for sensitivity, while specificity again showed homogeneity. The Spearman correlation coefficient for sensitivity and specificity was –0.94, which was judged to be sufficient to estimate a summary ROC curve. The curve in Figure 15 indicates that OVVOL volume has no clear accuracy in the prediction of non-pregnancy in IVF patients, even if a very low threshold for abnormality of the test would be chosen.

Table 21. Characteristics of included studies on ovarian volume (OVVOL) (computerised search using the test specific keyword *ovarian volume*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Ovarian volume (mL)	Ultrasonography equipment
Syrop *et al.*	Yes	Yes	Cycle	< 2 foll. 18 mm.	Clinical	Total < 8.6 mL, smallest < 3 mL	General Elect. 3600: 5 MHz probe
Lass *et al.*	Yes	Yes	Cycle	< 3 foll. 17 mm.	Clinical	MOV < 3 mL	Kretz Comb. 410: 5-7.5 Mhz probe
Sharara *et al.*	Yes	Yes	Cycle	Poor foll. development	Not applicable	MOV < 3 mL	Performa: 6.5 MHz probe
Schild *et al.*	Yes	Yes	Cycle	Not stated	Biochemical	MOV < 3 mL	Voluson 530D : 7.5 MHz probe
Bancsi *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18mm	Clinical	Total < 7 mL or < 8.6 mL	Toshiba SSA: 7.5 MHZ probe
Kupesic *et al.*	Yes	Yes	Cycle	Not applicable	Biochemical	Total < 7 mL	Combison 530 D: 7.5 Mhz probe
Järvelä *et al.*	Yes	Yes	Cycle	< 4 foll.	Clinical	MOV < 7 mL or < 3 mL	Kretz Comb 530
Fiçicioğlu *et al.*						Not stated	
Erdem *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm. or < 5 oocytes (MII)	Clinical	MOV < 2.98 mL	Aloka SSD1000: 5 MHz probe
Frattarelli *et al.*	Yes	Yes	Cycle	< 3 foll.	Biochemical	MOV < 2 mL or < 3 mL	Acuson 128: 7 MHz probe

Note: MOV = mean ovarian volume

Table 24. The occurrence of the ovarian volume test results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			**Prediction of non-pregnancy (pre-test probability = 80%)**		
LR range	Occurrence of test results in this range (%)	Posttest probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Posttest probability of non-pregnancy (%)
0-1	54	<20	0-1	68	<80
1-2	16	20-33	1-2	31	80-89
2-3	30	33-43	2-3	3	89-93
3-4	0	43-50	3-4	0	93-94
4-5	0	50-56	4-5	0	94-95
5-6	0	56-60	5-6	0	95-96
6-7	0	60-64	6-7	0	96-96.5
7-8	0	64-67	7-8	0	96.5-97
>8	0	>67	>8	0	>97

Table 22. Performance of the ovarian volume in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal volume result.

Author	Cycles (n)	Volume cut-off value (mL)	Prediction of poor response				Pre-volume probability (%)	Post-volume probability (%)	Proportion of patients/ cycles with abnormal v olume (%)
			Sens*	Spec†	LR+	DOR			
Syrop *et al.*	188	< 8.6	0.25	0.86	1.78	2.0	13	21	15
		< 3	0.17	0.91	1.95	2.1	13	22	10
Lass *et al.*	140	< 3	0.45	0.93	6.75	11.5	14	53	12
Sharara *et al.*	73	< 3	0.80	0.72	2.86	10.3	7	17	32
Schild *et al.*	152	< 3	0.11	0.90	1.10	1.1	18	20	10
Bancsi *et al.*	120	< 8.6	0.61	0.73	2.23	4.2	30	49	38
		< 7	0.39	0.85	2.51	3.5	30	52	23
Kupesic *et al.*	56	< 7	0.86	0.87	6.49	39.4	12	46	22
Järvelä *et al.*	60	< 3	0.08	0.94	1.30	1.3	18	25	6
		< 7	0.55	0.67	1.67	2.5	18	27	37
Fiçicioğlu *et al.*	58	< 4.9	0.73	0.53	1.50	2.7	43	53	59
Erdem *et al.*	32	< 2.98	0.75	0.81	4.00	13.0	50	80	47
Frattarelli *et al.*	267	< 2	0.17	0.94	2.83	3.2	9	21	7
		< 3	0.35	0.82	1.89	1.4	9	15	20

Note:if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 23. Performance of the ovarian volume in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal volume result.

Author	Cycles (n)	volume cut-off value (mL)	Prediction of non-pregnancy Sens*	Spec†	LR+	DOR	Pre-volume probability (%)	Post-volume probability (%)	Proportion of patients/ cycles with abnormal volume (%)
Syrop *et al.*	188	< 8.6	0.17	0.87	1.23	1.3	65	69	15
		< 3	0.11	0.93	1.44	1.5	65	72	10
Lass *et al*.	140	< 3	0.12	0.88	0.97	0.96	89	88	12
Schild *et al.*	152	< 3	0.12	0.97	3.60	3.9	80	93	10
Bancsi *et al*.	120	< 8.6	0.47	0.71	1.58	2.1	68	77	41
		< 7	0.27	0.79	1.33	1.5	68	74	25
Kupesic *et al.*	56	< 7	0.33	0.96	8.00	11.5	60	92	22
Järvelä *et al.*	60	< 3	0.08	0.96	1.80	1.9	63	75	6
		< 7	0.42	0.73	1.54	1.9	63	73	37
Erdem *et al.*	32	< 2.98	0.70	0.92	8.40	25.7	63	93	47
Frattarelli *et al.*	267	< 2	0.10	0.96	2.46	2.6	47	68	7
		< 3	0.22	0.82	1.27	1.4	47	53	20

*Note:*if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 14. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of **OVVOL** (ovarian volume) in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

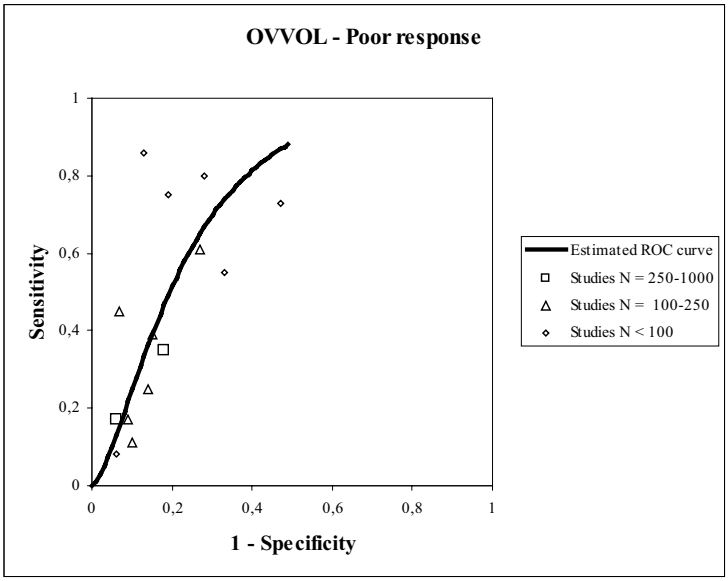
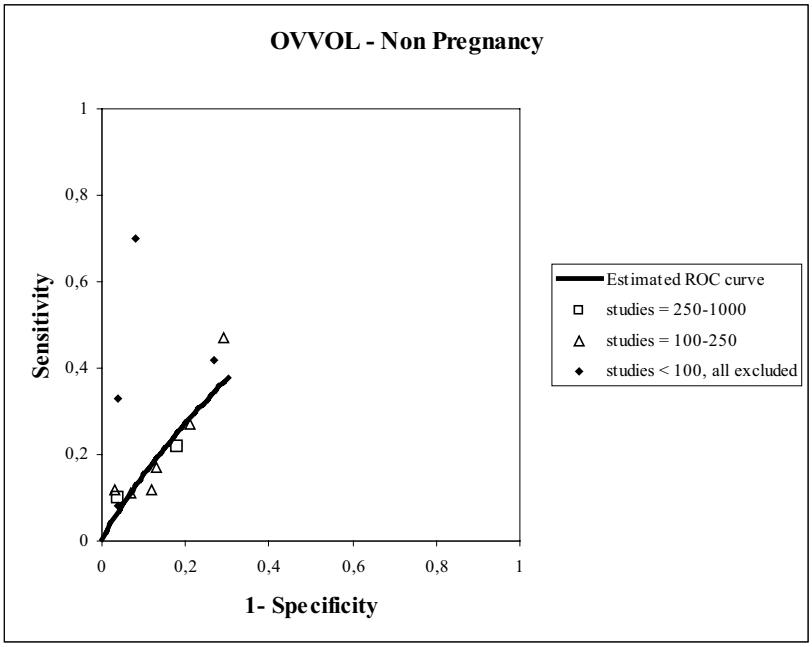


Figure 15. Estimated ROC curve and sensitivity/specificity points for studies reporting on the performance of **OVVOL** (ovarian volume) in the prediction of **non pregnancy,** after exclusion of 3 studies with verification bias. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.



Clinical value
Based on the estimated ROC curves in Figure 14 the probability of obtaining a certain test result for the OVVOL measurement is shown in Table 24 within a corresponding range of LRs for the prediction of poor ovarian response. Only at modest LRs, the post-test probability of poor response may approach 50%, while abnormal test results will be obtained in some 30% of tested cases. However, applying a more adequate positive likelihood level will result in virtually no cases being identified by the test. For non-pregnancy prediction, Table 25 shows that for higher LRs (>4), the post-test probability of non-pregnancy may increase to ~93–97%, assuming a pre-test probability of 80%. However, the probability that a test result will be in that range is close to zero. As false positive test results for both ovarian response and non-pregnancy prediction are not acceptable if patients are refused treatment, all this implies that the OVVOL is hardly suitable as a routine test for ovarian reserve assessment.

Ovarian vascular flow

Systematic review
Through the search we detected seven studies reporting on the predictive capacity of ovarian vascular flow parameters for ovarian response and/or the occurrence of pregnancy (Zaidi *et al.*, 1996, Engmann *et al.*, 1999a, Engmann *et al.*, 1999b, Kim *et al.*, 2002, Kupesic and Kurjak, 2002, Kupesic *et al.*, 2003, Popovic-Todorovic *et al.*, 2003b). In these studies ovarian flow was assessed either on cycle day 3 or after achievement of pituitary suppression with a GnRh agonist and before the onset of ovarian stimulation. As only the 2003 study by Kupesic (Kupesic *et al.*, 2003) could be included on a 2 x 2 for cross classification of the test result and the occurrence of poor response or non-pregnancy, it was not possible to carry out a formal meta-analysis (see addendum Table 25 and 26). Also, the studies used very different flow-derived predictors. Peak systolic velocity was used as the main predictor (Kupesic *et al.*, 2003). Others used ovarian stromal blood flow obtained by 3D power Doppler (Engmann *et al.*, 1999a).

Ovarian biopsy

Ovarian reserve depends on the number of primordial follicles in the ovarian cortex, which suggests that the obvious way to obtain an estimate would be to measure follicular density in an ovarian biopsy (Lass, 2001, Lass, 2004). Attempts were made to quantify the number of small antral follicles in small shallow biopsies taken during diagnostic laparoscopy from infertility patients (Lass *et al.*, 1997a) and there was a clear age-dependent decline in follicular density. Women over 35 years of age had only 30% of the quantities present in younger women. The number of follicles per unit of volume found in the biopsies was used to estimate the total and it was suggested that it could as such be potentially applied at the individual level. It was recognized though that the biopsy follicle density would not accurately represent the density in the whole ovary (Lass, 2001) and this seems indeed the case. Recently, several investigators have shown that follicle density varied greatly in small pieces of cortex, rendering information from biopsies as completely unreliable for an individual ovarian follicle content irrespective of how many were taken, their size and the location (Qu *et al.*, 2000, Schmidt *et al.*, 2003, Lambalk *et al.*, 2004, Sharara and Scott, 2004).

Table 25. Characteristics of included studies on ovarian stromal blood flow (OSF) (computerised search using the test specific keyword ovarian *ovarian stromal blood flow*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	OSF parameter	Ultrasonography equipment
Kupesic *et al.*	Yes	Yes	Cycle	Not applicable	Biochemical	Peak systolic velocity	Combison 530 D, 7.5 Mhz probe

Table 26. Performance of ovarian stromal flow (OSF) in the prediction of **non-pregnancy** in **IVF patients** and shift from pre-test to post-test probability of pregnancy for patients with an abnormal (= higher than the cut-off) ovarian stromal flow result.

Author	Cycles (n)	OSF cut-off value (flow index)	Prediction of non-pregnancy Sens*	Spec†	LR+	DOR	Pre-OSF probability (%)	Post-OSF probability (%)	Proportion of patients/cycles with abnormal OSF (%)
Kupesic et al	56	< 11	0.31	0.96	7.7	4.1	60	92	20
		≤ 13	0.85	0.23	1.1	1.5	60	64	82

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio

This indicates that the technique which is invasive and potentially harmful in terms of risks of adhesions and other complications of the surgical procedure is intrinsically unreliable and should therefore not be used to evaluate individual ovarian reserve. It is probably useful for research purposes to determine follicle density statistics in patient groups provided that group sizes are such that they compensate for the inherent extreme inter-biopsy and inter-individual spread of information (Qu *et al.*, 2000, Schmidt *et al.*, 2003, Webber *et al.*, 2003, Lambalk *et al.*, 2004). Finally, in the context of the current systematic review, there are no studies published that have evaluated ovarian biopsy follicle density for prediction of IVF outcome in terms of ovarian response and pregnancy rates.

Clomiphene Citrate Challenge Test

Systematic review

The computerized MEDLINE search detected 12 studies on the capacity of the Clomiphene Citrate Challenge Test (CCCT) to predict poor ovarian response and/or pregnancy after IVF (Tanbo *et al.*, 1989, Loumaye *et al.*, 1990, Tanbo *et al.*, 1990, Tanbo *et al.*, 1992, Csemiczky *et al.*, 1996, Kahraman *et al.*, 1997, van der Stege and van der Linden, 2001, Csemiczky *et al.*, 2002, Kwee *et al.*, 2003, Yanushpolsky *et al.*, 2003, Erdem *et al.*, 2004, Hendriks *et al.*, 2005a). Study characteristics of the included studies are listed in addendum Table 27. This table shows that many studies suffered from various sources of potential bias, especially selection bias. Also, definitions applied for poor ovarian response and for an abnormal CCCT result (based on either day-10 FSH alone or on both basal FSH and day-10 FSH results) varied considerably. Logistic regression analysis indicated that none of the study characteristics had a statistically significant impact on the reported predictive performance of the CCCT, neither for the outcome response nor for the outcome non-pregnancy. As a consequence, all studies were taken together for further analysis.

Accuracy of poor response prediction

For the prediction of ovarian response, sensitivities and specificities of each study are summarized in Table 28. Homogeneity could not be rejected for sensitivity (χ^2-test statistic: *P*-value 0.09), but had to be rejected for specificity (χ^2-test statistic: *P*-value <0.001). Therefore, calculation of one summary point estimate for sensitivity and specificity was not feasible. Moreover, values of the DOR (range 2.4–38.8) from the various studies appeared heterogeneous, indicating that the individual ROC curves were quite heterogeneous. Also, the Spearman correlation coefficient for sensitivity and specificity values was –0.46, which was judged not to be sufficient to estimate a summary ROC-curve. A plot of the sensitivity–specificity points in an ROC space is shown in Figure 16, showing the considerable heterogeneity which appeared not be attributable to differences in threshold level used.

Accuracy of non-pregnancy prediction

For the prediction on non-pregnancy, the sensitivities and specificities of each study are summarized in Table 29. Homogeneity was rejected for both sensitivity and specificity (χ^2-test statistic: *P*-value <0.001 and 0.04, respectively) and calculation of one summary point estimate for sensitivity and specificity was not meaningful. Also, the values of the DOR in the various studies (range 1.0–35.4) appeared non-homogeneous. A plot of sensitivity–specificity

points in an ROC space is shown in Figure 17. The Spearman correlation between sensitivity and specificity was –0.20, which again was judged not to be sufficient to estimate a summary ROC curve.

Clinical value

Because of the absence of estimated ROC curves for response and non-pregnancy prediction, the interrelation between positive LR, post-test probability and percentage of abnormal tests could not be calculated. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at 75 and 85%, respectively, it can be concluded from Figure 16 that hardly any study will fulfil these criteria. Moreover, in comparative studies the clinical performance of the CCCT in response prediction appeared not better than that of the AFC or FSH (Jain *et al.*, 2004, Hendriks *et al.*, 2005c). Regarding prediction of non-pregnancy, desired levels for a test that excludes cases from entering an IVF program should arbitrarily be set at 40% for sensitivity and 95% for specificity. The vast majority of studies fail to reach both criteria as shown in Figure 17 As such the CCCT performs no better than other tests like the AFC or basal FSH, especially because of a loss in specificity.

Figure 16. Sensitivity/specificity points for all studies reporting on the performance of the **CCCT** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points. Due to heterogeneity among studies no estimated summary ROC point of curve could be constructed. Reference lines indicate a desired level of test performance.

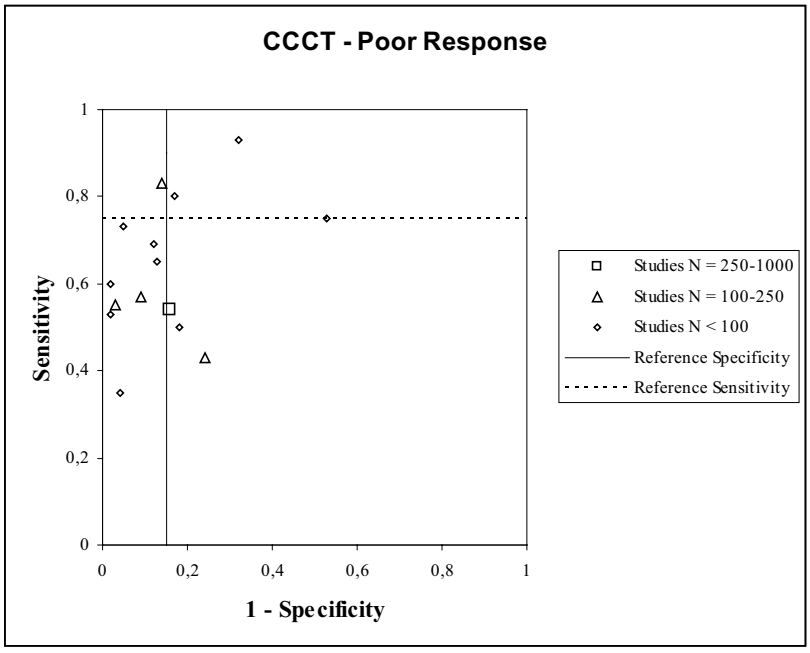


Table 27. Characteristics of included studies on the CCCT (computerised search using the test specific keyword *clomiphene citrate challenge test*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	FSH-assay
Tanbo *et al.*	No	Yes	Cycle	Cancel <3 foll.	Na	RIA: Amerlex
Tanbo *et al.*	No	No	Cycle	Cancel <2 foll.	Clinical	Fluoroimmunoassay: Delfia
Loumaye *et al.*	No	Yes	Cycle	Cancel <2 foll 20mm	Undefined	Immunoradiometr.: IRMA
Tanbo *et al.*	No	No	Cycle	Cancel <3 foll.	Ongoing	Fluoroimm.assay: Delfia
Csemiczky *et al.*	No	No	Cycle	Na	Clinical	RIA: Diagn Prod. Inc.
Kahraman *et al.*	No	Yes	Cycle	Undefined	Ongoing	Immunometr.: Diagn. Prod. Corp.
Vd Stege *et al.*	Yes	Yes	Cycle	Cancel <3 foll. 18mm	Clinical	RIA: Roche Diagn.
Csemiczky *et al.*	No	Yes	Cycle	Cancel <3 foll. 17mm	Ongoing	RIA: Farmos Group
Yanushpolsky *et al.*	Yes	No	Retrieval	Na	Delivery	Techn. Imm. Syst.: Bayer Corp.
Kwee *et al.*	Yes	Yes	Cycle	Poor response <6 oocytes	Na	Immunomet.: Amerlite/Delfia
Erdem *et al.*	Yes	Yes	Cycle	Cancel <4 foll. 15mm or Poor response <5 oocytes	Clinical	Chemolum. Immunometr. Assay
Hendriks *et al.*	Yes	Yes	Cycle	Poor response <4 oocytes or cancel no foll. growth	Ongoing	AxSYM immunoanal.: Abbott Lab.

Note: Na = not applicable.

Table 28. Performance of the CCCT in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal CCCT result.

Author	Cycles (n)	FSH cut-off value (IU/L)	Prediction of poor response				Pre-CCCT probability (%)	Post-CCCT probability (%)	Proportion of patients/ cycles with abnormal CCCT (%)
			Sens*	Spec†	LR+	DOR			
Tanbo *et al.*	109	Day7>26	0.55	0.97	17.7	37.8	40	92	24
Tanbo *et al.*	70	Day10>26	0.75	0.47	1.4	2.7	46	55	63
Loumaye *et al.*	114	Day3+10>26.03	0.83	0.86	6.0	31.0	5	25	18
Tanbo *et al.*	165	Day10>12	0.57	0.91	5.9	12.5	49	85	33
Kahraman *et al.*	198	Day10>10	0.43	0.76	1.8	2.4	25	37	29
Vd Stege *et al.*	51	Day3 or 10>10	0.50	0.82	2.7	4.4	4	10	20
Csemiczky *et al.*	279	Day10>10	0.54	0.84	3.3	6.1	25	53	26
Kwee *et al.*	56	Day3+10>14	0.93	0.68	2.9	30.1	27	52	48
		Day3+10>16	0.80	0.83	4.7	19.4	27	63	34
		Day3+10>18	0.73	0.95	15.0	53.6	27	85	23
		Day3+10>20	0.60	0.98	24.6	60.0	27	90	18
		Day3+10>22	0.53	0.98	21.9	45.7	27	89	16
Erdem *et al.*	32	Day3 or 10>10	0.69	0.88	5.5	15.4	50	85	41
Hendriks *et al.*	63	Day10>10	0.65	0.87	5.0	12.2	27	65	27
		Day10>15	0.35	0.96	8.1	12.0	27	75	13

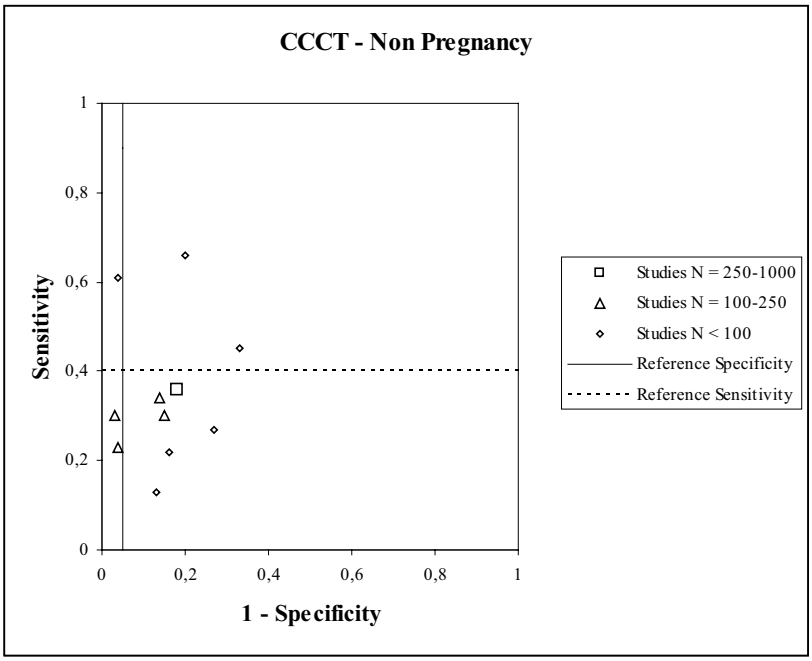
Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 29. Performance of the CCCT in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal CCCT result.

Author	Cycles (n)	FSH cut-off value (IU/L)	Prediction of non-pregnancy				Pre-CCCT probability (%)	Post-CCCT probability (%)	Proportion of patients/ cycles with abnormal CCCT (%)
			Sens*	Spec†	LR+	DOR			
Tanbo *et al.*	70	Day10>26	0.66	0.80	3.3	7.8	93	98	63
Loumaye *et al.*	114	Day3+10>26.03	0.23	0.96	6.5	8.2	76	95	19
Tanbo *et al.*	165	Day10>12	0.34	0.86	2.4	3.1	96	98	33
Csemiczky *et al.*	53	Day10>7	0.61	0.96	14.5	35.4	58	95	37
Kahraman *et al.*	198	Day10>10	0.30	0.85	2.4	3.0	92	96	29
Vd Stege *et al.*	51	Day3 or 10>10	0.22	0.84	1.4	1.5	63	70	20
Csemiczky *et al.*	140	Day10>10	0.30	0.97	8.6	11.8	79	97	24
Yanushpolsky *et al.*	483	Day10>10	0.36	0.82	2.0	2.5	62	76	29
Erdem *et al.*	32	Day3 or 10>10	0.45	0.67	1.4	1.6	63	69	41
Hendriks *et al.*	63	Day10>10	0.27	0.73	1.0	1.0	76	77	27
		Day10>15	0.13	0.87	0.9	0.9	76	75	13

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 17. Sensitivity/specificity points for all studies reporting on the performance of the **CCCT** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points. Due to heterogeneity among studies no estimated summary ROC point of curve could be constructed. Reference lines indicate a desired level of test performance.



Exogenous FSH Ovarian Reserve Test

Systematic review

We detected three studies from the literature reporting on the predictive capacity of the exogenous FSH ORT (EFORT) that were suitable for data extraction (Fanchin *et al.*, 1994, Kwee *et al.*, 2003, Yong *et al.*, 2003). The characteristics of these studies are listed in addendum Table 30.

Accuracy of poor response prediction

The individual values for sensitivity and specificity pairs are summarized in Table 31 and plotted in Figure 18. As can be seen from this ROC space, the three detected studies report sensitivities around 80%, whereas specificities vary around 60% in the study of Kwee *et al*. and Yong *et al.* and above 90% in the study of Fanchin *et al*. In view of these different results between the studies, further assessment of heterogeneity appeared not useful and therefore a summary point or curve in the ROC space could not be constructed.

Accuracy of non-pregnancy prediction

No single study reported on the predictive accuracy using the outcome pregnancy as test reference.

Clinical value
Because of the absence of an estimated ROC curve for poor response prediction, the interrelation between positive LR, post-test probability and percentage of abnormal tests could not be calculated. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at a minimum level of 75 and 85%, respectively, it can be concluded from Figure 18 that only one study fulfils these criteria (Fanchin *et al.*, 1994). Especially, the false positive prediction may hamper the use of this test if a high level of detection is needed and patients are refused IVF on the basis of the test result. Finally, in comparison to basal tests, challenge tests should clearly improve prediction if they are to be preferred.

Gonadotrophin: releasing hormone agonist stimulation test

Systematic review
Through the search and selection strategy, a total of four studies reporting on the predictive capacity of the Gonadotrophin releasing hormone agonist stimulation test (GAST) were identified and considered suitable for data extraction and meta-analysis (Ranieri *et al.*, 1998, Padilla *et al.*, 1990, Winslow *et al.*, 1991, Hendriks *et al.*, 2005b). Characteristics of the included studies are listed in addendum Table 32. Considerable variation among the definitions of poor response and the study quality and design characteristics was observed, but as only three studies reported on each of the two endpoints, a systematic analysis of these study characteristics was not indicated.

Accuracy of poor response prediction
There were three studies that reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table 33. Calculation of one summary point estimate for sensitivity and specificity was not meaningful as both test characteristics shown in Figure 19 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.014). As the Spearman correlation coefficient for sensitivity and specificity was –0.57, it appeared justified to estimate a summary ROC curve as shown in Figure 19.

Accuracy of non-pregnancy prediction
There were also three studies that reported on the capacity of the GAST to predict non-pregnancy after IVF. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table 34. Again, sensitivity and specificity, as shown in Figure 20, were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.005). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.98, sufficient to estimate a summary ROC curve (Figure 20).

Table 30. Characteristics of included studies on the EFORT (computerised search using the test specific keyword *EFORT*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Estradiol-assay
Fanchin *et al.*	Yes	Yes	Cycle	< 3 oocytes	na	Estradiol-60 Amerlite (Kodak clin. Diagn. UK)
Kwee *et al.*	Yes	Yes	Cycle	< 6 oocytes	na	Amerlite (Amersham UK)
Yong *et al.*	No	Yes	Cycle	< 4 oocytes or cancel	na	Radioimmunoassay

Note: Na = not applicable.

Table 31. Performance of the EFORT in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal EFORT result.

Author	Cycles (n)	Estradiol cut-off value (pmol/l)	Prediction of poor response Sens*	Spec†	LR+	DOR	Pre-EFORT probability (%)	Post-EFORT probability (%)	Proportion of patients/ cycles with abnormal EFORT (%)
Fanchin *et al.*	52	< 110	0.79	0.92	2.7	42.8	27	79	27
Kwee *et al.*	54	< 110	0.64	0.68	1.98	3.7	26	41	41
		< 120	0.64	0.65	1.8	3.3	26	39	43
		< 130	0.71	0.65	2.0	4.6	26	42	44
		< 140	0.79	0.60	1.96	5.5	26	41	50
		< 150	0.86	0.58	2.0	8.1	26	41	54
Yong *et al.*	46	< 124	0.50	0.68	1.6	2.2	17	25	35

*Note:*if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 18. Sensitivity/specificity points for all studies reporting on the performance of the **EFORT** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points. Due to heterogeneity among studies no estimated summary ROC point of curve could be constructed. Reference lines indicate a desired level of test performance.

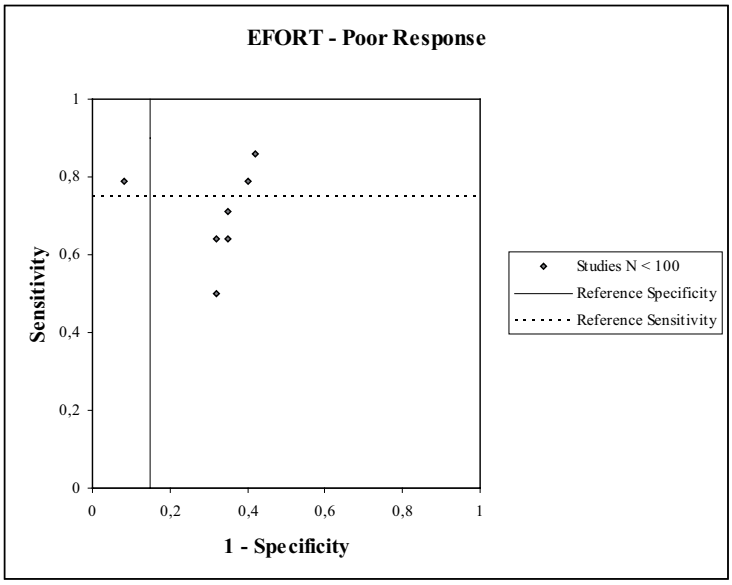


Figure 19. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of the **GAST** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.

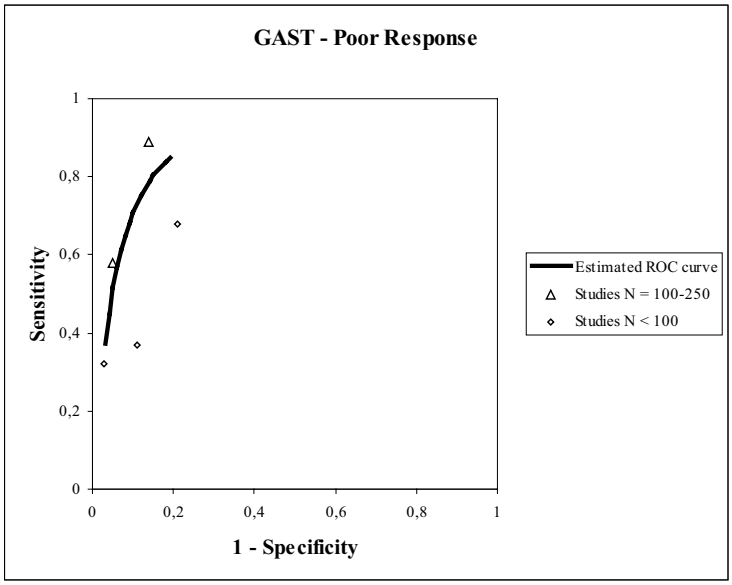


Table 32. Characteristics of included studies on GAST (computerised search using the test specific keyword *gonadotrophin agonist stimulation test*).

Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy	Estradiol assay
Padilla *et al.*	No	No	Cycle	Not applicable	Clinical	RIA (Diagnostic Products USA)
Winslow *et al.*	Yes	Yes	Cycle	Not stated	Clinical	Radioimmunoassay (Pantex CA)
Ranieri *et al.*	No	Yes	Cycle	< 5 foll. 15 mm.	Not applicable	RIA (Amersham Int. UK)
Hendriks *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18 mm.	ongoing	AxSYM immunoanalyser (Abbott Lab USA)

Table 33. Performance of GAST in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal GAST result.

Author	Cycles (n)	Estradiol cut-off value (pmol/l)	Prediction of poor response Sens*	Spec†	LR+	DOR	Pre-GAST probability (%)	Post-GAST probability (%)	Proportion of patients/ cycles with abnormal GAST (%)
Winslow *et al.*	228	$\Delta E_2 < ?$	0.58	0.95	11.5	26.1	5	39	8
Ranieri *et al.*	177	$\Delta E_2 < 180$	0.89	0.86	6.4	53.0	27	70	34
Hendriks *et al.*	57	$\Delta E_2 < 80$	0.32	0.97	12.0	17.1	33	86	12
		$\Delta E_2 < 100$	0.37	0.89	3.5	4.6	33	64	19
		$\Delta E_2 < 180$	0.68	0.79	3.3	8.1	33	62	37

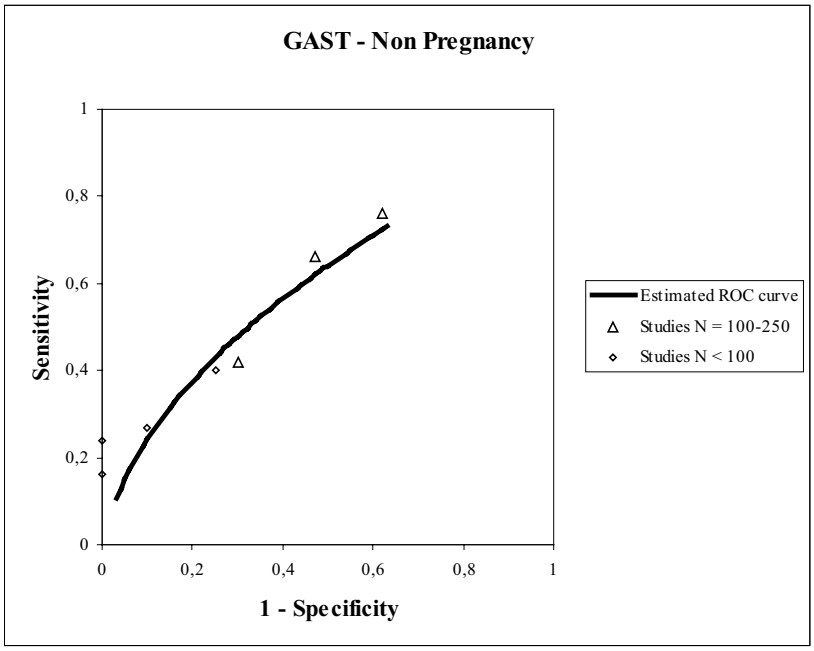
Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Table 34. Performance of GAST in the prediction of **non-pregnancy** in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal GAST result.

Author	Cycles (n)	Estradiol cut-off value (pmol/l)	Prediction of non-pregnancy Sens*	Spec†	LR+	DOR	Pre-GAST probability (%)	Post-GAST probability (%)	Proportion of patients/ cycles with abnormal GAST (%)
Padilla *et al.*	97	$\Delta E_2 < ?$	0.27	0.90	2.8	3.5	68	86	22
Winslow *et al.*	228	$\Delta E_2 < 50$	0.42	0.70	1.4	1.69	77	82	39
		$\Delta E_2 < 75$	0.66	0.53	1.4	2.2	77	82	62
		$\Delta E_2 < 100$	0.76	0.38	1.2	1.92	77	80	73
Hendriks *et al.*	57	$\Delta E_2 < 80$	0.16	1.00	2.4	2.7	79	89	12
		$\Delta E_2 < 100$	0.24	1.00	3.6	4.6	79	92	19
		$\Delta E_2 < 180$	0.40	0.75	1.6	2.0	79	86	37

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio.

Figure 20. Estimated ROC curve and sensitivity/specificity points for all studies reporting on the performance of the **GAST** in the prediction of **non pregnancy**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points.



Clinical value

Based on the summary ROC curves depicted in Figure 19, a range of positive LRs was calculated and for each ratio, pre-GAST-test probabilities of poor response or non-pregnancy (set at 20 and 80%, respectively) were converted into post-GAST-test probabilities. Table 35 depicts the probability of obtaining a certain GAST test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a modest LR of 4–5, the post-GAST-test probability of poor response will not be higher than ~50%, while the chance of obtaining such a test result is quite high, 49%. However, only with an extreme threshold a post-test probability of poor response that approaches 70% can be retained in a considerable number of cases (30%).

For prediction of non-pregnancy, extreme threshold levels are necessary to obtain a modest positive LR of 4–5, leading to a post-test pregnancy rate of approximately 5%. Such abnormal test results occur only in a very limited number of patients, while the false positive rate will lead to unnecessary exclusions from IVF programs if the test is used in a diagnostic fashion. It can be concluded that with the use of the GAST in regularly cycling women, the accuracy in the prediction of poor response is quite high and could match with those obtained by the use of the AFC. For non-pregnancy prediction the test may only be adequate at a very low threshold level, where hardly any abnormal tests can be found. The results show that the GAST is a candidate for more extensive confirmation research.

Multivariate models

Systematic review

Through the search and selection strategy, a total of nine studies reporting on the predictive capacity of several multivariate models were identified and considered suitable for data extraction and meta-analysis (Balasch *et al.*, 1996, Ranieri *et al.*, 1998, Creus *et al.*, 2000, Fabregues *et al.*, 2000, Bancsi *et al.*, 2002a, van Rooij *et al.*, 2002, Durmusoglu *et al.*, 2004, Erdem *et al.*, 2004, Muttukrishna *et al.*, 2004). Characteristics of the included studies are listed in addendum Table 36. As with most studies on ORTs, definitions for poor response varied considerably. It should be noted that none of the multifactor studies revealed usable data on pregnancy prediction. Moreover, the total number of cases included in these aggregated studies is modest (n=991).

Accuracy of poor response prediction

All ten studies only reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response are summarized in Table 37. Calculation of one summary point estimate for sensitivity and specificity did not appear to be possible, as both test characteristics (shown in Figure 21) were heterogeneous among studies (χ^2-test statistic: P-value for sensitivity <0.001 and P-value for specificity 0.014). As the Spearman correlation coefficient for sensitivity and specificity was –0.45, it appeared unjustified to estimate a summary ROC curve. Regression analysis showed that the performance of one particular test model was not superior to the other, as can also be seen in addendum Table 36 from the listing of sensitivities and specificities.

Clinical value

The impossibility of creating summary characteristics makes it difficult to assess the interrelation between positive LR, post-test probability and percentage of abnormal tests. Obviously, clinical value can only be discussed regarding prediction of poor response. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at 75 and 85%, respectively, it can be concluded from Figure 21 that only one study will fulfil these criteria (Bancsi *et al.*, 2002a). Especially, the false positive prediction may hamper the use of this test if a high level of detection is needed and patients are refused IVF on the basis of this test. From these data it seems that compared to other ORTs, multifactor models do not create a definite improvement in predictive capacity.

Table 35. The occurrence of the GAST volume results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%.

Prediction of poor response (pre-test probability = 20%)			**Prediction of non-pregnancy (pre-test probability = 80%)**		
LR range	Occurrence of test results in this range (%)	Posttest probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Posttest probability of non-pregnancy (%)
0-1	31	<20	0-1	70	<80
1-2	8	20-33	1-2	22	80-89
2-3	5	33-43	2-3	2	89-93
3-4	6	43-50	3-4	6	93-94
4-5	3	50-56	4-5	0	94-95
5-6	4	56-60	5-6	0	95-96
6-7	5	60-64	6-7	0	96-96.5
7-8	7	64-67	7-8	0	96.5-97
>8	30	>67	>8	0	>97

Table 36. Characteristics of included studies on multi-variate models (computerised search using the test specific keywords multifactor, multivariate, prediction model and logistic model).

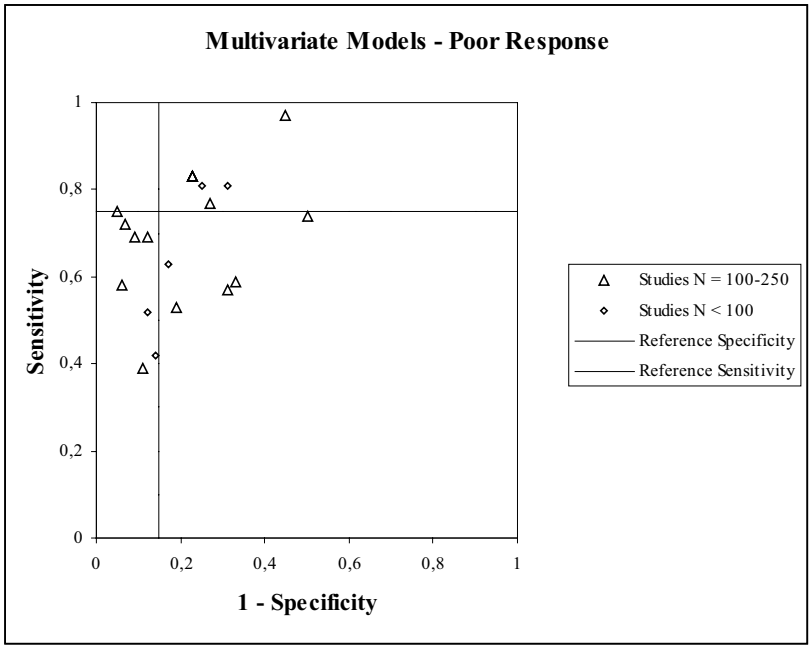
Author	Consecutive	One cycle per couple	Data per	Definition Poor response/Cancel	Definition Pregnancy
Balasch *et al.*	Yes	Yes	Cycle	< 2 foll. 17 mm. or < 5 foll. 14 mm.	Not applicable
Fabregues *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable
Ranieri *et al.*	No	Yes	Cycle	< 5 foll. 15 mm.	Not applicable
Creus *et al.*	Yes	Yes	Cycle	< 3 foll. 14 mm.	Not applicable
Bancsi *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll. 18 mm.	Not applicable
Van Rooij *et al.*	Yes	Yes	Cycle	< 4 oocytes or < 3 foll.	Not applicable
Muttukrishna *et al.*	No	Yes	Cycle	< 4 foll. 15 mm.	Not applicable
Erdem *et al.*	Yes	Yes	Cycle	Cancel <4 foll. 15mm or Poor response <5 oocytes	Not applicable
Durmusoglu *et al.*	No	No	Cycle	Poor foll. growth or < 3 oocytes (MII)	Not applicable

Table 37. Performance of multi-variate models in the prediction of **poor response** in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal test result.

Author	Cycles (n)	Test Model	Prediction of poor response				Pre-test probability (%)	Post-test probability (%)	Proportion of patients/ cycles with abnormal test (%)
			Sens*	Spec†	LR+	DOR			
Balash *et al.*	120	Age + FSH	0.53	0.81	2.8	4.8	33	58	30
		Age + inhibin B	0.59	0.67	1.8	2.8	33	48	42
		Inhibin B + FSH	0.57	0.69	1.8	3.0	33	48	40
		Age + FSH + inhibin	0.39	0.89	3.5	5.2	33	64	21
Fabregues *et al.*	80	FSH + Inhibin B	0.42	0.86	3.0	4.4	35	63	24
Ranieri *et al.*	177	FSH + GAST	0.97	0.55	2.2	39.5	33	45	59
Creus *et al.*	120	Age + FSH	0.83	0.77	3.6	16.3	33	65	43
		Age + inhibin B	0.74	0.50	1.5	2.8	33	43	58
		FSH + inhibin B	0.77	0.73	2.9	9.1	33	58	44
		Age + FSH + inhibin	0.83	0.77	3.6	16.3	33	65	43
Bancsi *et al.*	120	FSH + inhibin B	0.58	0.94	9.7	21.6	30	81	22
		AFC + inhibin B	0.69	0.88	5.6	16.3	30	71	29
		AFC + FSH	0.72	0.93	10.3	34.2	30	81	27
		AFC + inihbin B +	0.75	0.95	15	57.0	30	87	26
Van Rooij *et al.*	119	AMH + inhibin B	0.69	0.91	7.7	22.5	29	75	27
Muttukrishna *et al.*	69	FSH + inhibin B +	0.63	0.83	3.7	8.3	25	65	29
Erdem *et al.*	32	CCCT + age	0.81	0.69	2.6	9.5	50	72	56
		CCCT + age + OVVOL	0.81	0.75	3.2	12.8	50	76	53
Durmusoglu *et al.*	91	Age + AFC	0.52	0.88	4.3	7.9	26	62	23

Note: if a study reported on multiple cut-off values, data for all cut-off values are shown. *sensitivity, †specificity, LR+ = likelihood ratio for a positive testresult. DOR = diagnostic odds ratio. OV = ovarian volume, CCCT = clomiphene citrate challenge test, AFC = antral follicle count, GAST = GnRH agonist stimulation test.

Figure 21. Sensitivity/specificity points for all studies reporting on the performance of **multivariate models** in the prediction of **poor response**. Studies reporting on several cut off points are represented by an equivalent number of sens/spec points. Due to heterogeneity among studies no estimated summary ROC point of curve could be constructed. Reference lines indicate a desired level of test performance.



Implications for daily practice

Wth the postponement of childbearing, the age-related fertility decline has been shown to play an important role in the increase in infertility among couples who are trying to conceive. In IVF treatment, this age effect has been shown in much accumulated data. Because of the variation of female fertility within a certain age category, the need was felt for tests which better identified cases with a state of ovarian reserve that is clearly too low for their age. Because a benchmark for ovarian reserve status in the sense of quantity and quality is lacking, the occurrence of poor ovarian response to maximal stimulation and the occurrence of pregnancy in IVF are used as parameters to assess the accuracy of the test. The ideal ORT should identify a substantial percentage of IVF-indicated cases which have a practically zero chance of becoming pregnant because of the adverse effects of diminished ovarian reserve in a series of treatment cycles. Those cases can be refrained from entering the programme, as the very high costs involved will have only minimal results. If not too expensive and not too demanding for the patient, such a test would be readily embraced by physicians, patients, health politicians and insurance companies. From the systematic and meta-analytic reviews presented here, it can be concluded that the ORTs known to date have only very modest predictive properties and are therefore far from suitable for relevant clinical use. Although mostly cheap and not very demanding, their accuracy, especially in the prediction of the occurrence of pregnancy, is

very limited. If a high threshold is used, to prevent couples from wrongly being refused IVF, a very small minority of IVF-indicated cases (~3%) were identified as having unfavourable prospects in an IVF treatment cycle (pregnancy rate for that cycle = 5%). It should be noted that the use of pregnancy as outcome parameter for the assessment of ovarian reserve status may be insufficient if only one exposure cycle is taken into account. As such, the possibility of misjudgement on the basis of currently known ORTs is hard to rule out. This implies that the use of the test as a method to deny treatment to assumed ovarian aged women should be declined and, as a consequence the test should not be applied on a regular basis and should only be used for counselling or screening purposes. Accuracy of testing for the occurrence of poor ovarian response to stimulation appeared to be clearly better than for the occurrence of pregnancy. This may be understood in the light of the following factors: (i) that the chance of pregnancy after IVF depends on many more factors than ovarian reserve alone, (ii) that the occurrence of pregnancy after an ORT was usually evaluated in only one IVF cycle and as such may not accurately represent a female's true reproductive capacity and (iii) that the response to stimulation is likely to represent the size of the cohort of FSH-sensitive follicles continuously present in the ovaries and is directly related to the magnitude of ovarian reserve (i.e. the remaining primordial follicle pool (Gougeon, 1984). Poor ovarian response has been associated with a reduced chance of pregnancy in the actual treatment cycle as well as in subsequent cycles and as such may well be indicative of ovarian reserve status in both the quantitative and qualitative sense (Ulug *et al.*, 2003, Klinkert *et al.*, 2004, Klinkert *et al.*, 2005a). Accurate prediction of poor response could therefore have clinical value if the pregnancy prospects are so unfavourable that a predicted poor responder would be denied treatment. Accuracy in response prediction, however, will only be high if the false positives are prevented by using extreme threshold levels, implicating that only minor percentages of abnormal tests will be found and many future poor responders will pass unrecognized. At the same time it is necessary to know whether the predicted poor responder indeed has very low prospects for success in subsequent cycles. As much of this is unknown at the present time, the use of any ORT for poor response prediction cannot be supported, not even if it would be used for adapting the treatment schedule in anticipated poor responders, as an altered treatment schedule has consistently been shown to be effective in women with a severely reduced size of follicle cohort (Tarlatzis *et al.*, 2003, Klinkert, 2005, Klinkert *et al.*, 2005a). One aspect of clinical value deserves some special attention. ORTs are mostly used as a diagnostic test, indicating that in case of an abnormal test result, the diagnosis that there is diminished ovarian reserve is made (Scott and Hofmann, 1995, Levi *et al.*, 2001). From the fact that for evaluation of the test, proxy variables of true ovarian reserve (poor ovarian response and non-pregnancy) are used and that false positive test results may eliminate couples from the IVF trail even if they do have adequate prospects, it becomes clear that ORTs may better be considered as *screening tests*. All this implies that an abnormal test necessitates confirmation by another test. This other test may, for instance, be a first IVF attempt where ovarian response is the additional test. Alternatively, combinations of independent predictive tests or repeating of the initial test could improve the diagnostic performance of the single test (Ng *et al.*, 2000, Bancsi *et al.*, 2002b, van Rooij *et al.*, 2002, Popovic-Todorovic *et al.*, 2003a, Popovic-Todorovic *et al.*, 2003b, Bancsi *et al.*, 2004a, Bancsi *et al.*, 2004b). As poor ovarian response will provide some information on ovarian reserve status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy. Once a poor response is obtained, the question arises whether this finding is based on depleted ovaries or other causes, like underdosing

for instance, based on the presence of certain FSH receptor polymorphisms (Perez *et al.*, 2000, Behre *et al.*, 2005, Greb *et al.*, 2005, de Koning *et al.*, 2005). A repeat cycle with adequate, maximal stimulation or a *post hoc*-performed ORT [basal FSH or AFC (Hendriks *et al.*, 2005c)] may correctly classify the poor responder patient as having an aged ovary and may correctly suggest that they refrain from further treatment (Klinkert *et al.*, 2004). It should be remembered that the purpose of any ORT is the identification of women with poor ovarian reserve for their age. This implies that chronological age always is the first step in ovarian reserve assessment. In young women, ORTs may help to classify poor responders and in direct management in these cases by estimating the size of the FSH-sensitive cohort. In older women, ORTs may help to identify those cases that, in spite of their age, still may have acceptable chances of becoming pregnant through IVF as the quantity of response to stimulation is anticipated to be normal or even high (Klinkert *et al.*, 2005b). Future perspectives in this research field may be found in studies where success rates in cumulative treatment cycles or in units of time (1-year treatment periods) are analysed to answer the question of whether any test will correctly identify those couples who will not become pregnant in such series of exposures. Novel tests that most accurately estimate the age at which menopause is expected to take place in an individual woman may facilitate the estimation of the remaining reproductive potential at a certain age. Such tests will probably be based on family history (age at menopause of mother) or will comprise testing for genetic markers, which may be discovered from large-scale population genetic screening.

References

Abma JC, Chandra A, Mosher WD, Peterson LS and Piccinino LJ (1997) Fertility, family planning, and women's health: new data from the 1995 National Survey of Family Growth. Vital Health Stat 23,1–114.

Akande VA, Keay SD, Hunt LP, Mathur RS, Jenkins JM and Cahill DJ (2004) The practical implications of a raised serum FSH and age on the risk of IVF treatment cancellation because of a poor ovarian response. J Assist Reprod Genet 21,257–262.

Anonymous (1995) Pregnancies and births resulting from in vitro fertilization: French national registry analysis of data 1986 to 1990. FIVNAT (French In Vitro National). Fertil Steril 64,746–756.

Balasch J, Creus M, Fabregues F, Carmona F, Casamitjana R, Ascaso C and Vanrell JA (1996) Inhibin, follicle-stimulating hormone, and age as predictors of ovarian response in in vitro fertilization cycles stimulated with gonadotropin-releasing hormone agonist-gonadotropin treatment. Am J Obstet Gynecol 175,1226–1230.

Bancsi LF, Huijs AM, Den Ouden CT, Broekmans FJ, Looman CW, Blankenstein MA and te Velde ER (2000) Basal follicle-stimulating hormone levels are of limited value in predicting ongoing pregnancy rates after in vitro fertilization. Fertil Steril 73,552–557.

Bancsi LF, Broekmans FJ, Eijkemans MJ, de Jong FH, Habbema JD, te Velde ER (2002a) Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 77,328–336.

Bancsi LFJMM, Broekmans FJM, Eijkemans MJC, de Jong FH, Habbema JDF and te Velde ER (2002b) Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 77,328–336.

Bancsi LF, Broekmans FJ, Mol BW, Habbema JD and te Velde ER (2003) Performance of basal follicle-stimulating hormone in the prediction of poor ovarian response and failure to become pregnant after in vitro fertilization: a meta-analysis. Fertil Steril 79,1091–1100.

Bancsi LF, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2004a) Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. Fertil Steril 81,35–41.

Bancsi LF, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2004b) Predicting poor ovarian response in IVF: use of repeat basal FSH measurement. J Reprod Med 49,187–194.

Bassil S, Godin PA, Gillerot S, Verougstraete JC and Donnez J (1999) In vitro fertilization outcome according to age and follicle-stimulating hormone levels on cycle day 3. J Assist Reprod Genet 16,236–241.

Behre HM, Greb RR, Mempel A, Sonntag B, Kiesel L, Kaltwasser P, Seliger E, Ropke F, Gromoll J, Nieschlag E and Simoni M (2005) Significance of a common single nucleotide polymorphism in exon 10 of the follicle-stimulating hormone (FSH) receptor gene for the ovarian response to FSH: a pharmacogenetic approach to controlled ovarian hyperstimulation. Pharmacogenet Genomics 15,451–456.

Block E (1952) Quantitative morphological investigations of the follicular system in women. Variations at different ages. Acta Anat (Basel) 14 (Suppl 16),108–123.

de Boer EJ, den TI, te Velde ER, Burger CW, Klip H and van Leeuwen FE (2002) A low number of retrieved oocytes at in vitro fertilization treatment is predictive of early menopause. Fertil Steril 77,978–985.

de Bruin JP and te Velde ER (2004) Female reproductive ageing: concepts and consequences. In Tulandi T and Gosden RG (eds) *Preservation of Fertility*. London, UK: Taylor & Francis, p. 3.

Chae HD, Kim CH, Kang BM and Chang YS (2000) Clinical usefulness of basal FSH as a prognostic factor in patients undergoing intracytoplasmic sperm injection. J Obstet Gynaecol Res 26,55–60.

Chan YF, Ho PC, So WW and Yeung WS (1993) Basal serum pituitary hormone levels and outcome of in vitro fertilization utilizing a flare nasal gonadotropin releasing hormone agonist and fixed low-dose follicle-stimulating hormone/human menopausal gonadotropin regimen. J Assist Reprod Genet 10,251–254.

Chang MY, Chiang CH, Chiu TH, Hsieh TT, Soong YK (1998a). The antral follicle count predicts the outcome of pregnancy in a controlled ovarian hyperstimulation/intrauterine insemination program. J Assist Reprod Genet 15,12–17.

Chang MY, Chiang CH, Hsieh TT, Soong YK and Hsu KH (1998b) Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil Steril 69,505–510.

Chuang CC, Chen CD, Chao KH, Chen SU, Ho HN and Yang YS (2003) Age is a better predictor of pregnancy potential than basal follicle-stimulating hormone levels in women undergoing in vitro fertilization. Fertil Steril 79,63–68.

Collins JA, Burrows EA and Wilan AR (1995) The prognosis for live birth among untreated infertile couples. Fertil Steril 64,22–28.

Creus M, Penarrubia J, Fabregues F, Vidal E, Carmona F, Casamitjana R, Vanrell JA and Balasch J (2000) Day 3 serum inhibin B and FSH and age as predictors of assisted reproduction treatment outcome. Hum Reprod 15,2341–2346.

Csemiczky G, Wramsby H and Landgren BM (1996) Luteal phase oestradiol and progesterone levels are stronger predictors than follicular phase follicle stimulating hormone for the outcome of in-vitro fertilization treatment in women with tubal infertility. Hum Reprod 11,2396–2399.

Csemiczky G, Harlin J and Fried G (2002) Predictive power of clomiphene citrate challenge test for failure of in vitro fertilization treatment. Acta Obstet Gynecol Scand 81,954–961.

Deeks JJ (2001) Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 323,157–162.

Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA and Bezemer PD (2002) Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2,9.

Durmusoglu F, Elter K, Yoruk P and Erenus M (2004) Combining cycle day 7 follicle count with the basal antral follicle count improves the prediction of ovarian response. Fertil Steril 81,1073–1078.

Ebrahim A, Rienhardt G, Morris S, Kruger TF, Lombard CJ and Van der Merwe JP (1993) Follicle stimulating hormone levels on cycle day 3 predict ovulation stimulation response. J Assist Reprod Genet 10,130–136.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW and Habbema JD (1994). The prediction of the chance to conceive in subfertile couples. Fertil Steril 61,44–52.

Engmann L, Sladkevicius P, Agrawal R, Bekir J, Campbell S and Tan SL (1999a) The pattern of changes in ovarian stromal and uterine artery blood flow velocities during in vitro fertilization treatment and its relationship with outcome of the cycle. Ultrasound Obstet Gynecol 13,26–33.

Engmann L, Sladkevicius P, Agrawal R, Bekir JS, Campbell S and Tan SL (1999b) Value of ovarian stromal blood flow velocity measurement after pituitary suppression in the prediction of ovarian responsiveness and outcome of in vitro fertilization treatment. Fertil Steril 71,22–29.

Erdem A, Erdem M, Biberoglu K, Hayit O, Arslan M and Gursoy R (2002) Age-related changes in ovarian volume, antral follicle counts and basal FSH in women with normal reproductive health. J Reprod Med 47,835–839.

Erdem M, Erdem A, Gursoy R and Biberoglu K (2004) Comparison of basal and clomiphene citrate induced FSH and inhibin B, ovarian volume and antral follicle counts as ovarian reserve tests and predictors of poor ovarian response in IVF. J Assist Reprod Genet 21,37–45.

Esposito MA, Coutifaris C and Barnhart KT (2002) A moderately elevated day 3 FSH concentration has limited predictive value, especially in younger women. Hum Reprod 17,118–123.

Evers JL, Slaats P, Land JA, Dumoulin JC and Dunselman GA (1998) Elevated levels of basal estradiol-17beta predict poor response in patients with normal basal levels of follicle-stimulating hormone undergoing in vitro fertilization. Fertil Steril 69,1010–1014.

Fabregues F, Balasch J, Creus M, Carmona F, Puerto B, Quinto L, Casamitjana R and Vanrell JA (2000) Ovarian reserve test with human menopausal gonadotropin as a predictor of in vitro fertilization outcome. J Assist Reprod Genet 17,13–19.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A and Frydman R (1994) Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Hum Reprod 9,1607–1611.

Fasouliotis SJ, Simon A and Laufer N (2000) Evaluation and treatment of low responders in assisted reproductive technology: a challenge to meet. J Assist Reprod Genet 17,357–373.

Fiçicioqlu C, Kutlu T, Demirbasoglu S and Mulayim B (2003) The role of inhibin B as a basal determinant of ovarian reserve. Gynecol Endocrinol 17,287–293.

Fisch JD and Sher G (2002) The antral follicle count (AFC) correlates with the metaphase II oocytes and ART cycle outcome: an update. Fertil Steril 78,S90.

Frattarelli JL, Lauria-Costab DF, Miller BT, Bergh PA and Scott RT (2000) Basal antral follicle number and mean ovarian diameter predict cycle cancellation and ovarian responsiveness in assisted reproductive technology cycles. Fertil Steril 74,512–517.

Frattarelli JL, Levi AJ, Miller BT and Segars JH (2003) A prospective assessment of the predictive value of basal antral follicles in in vitro fertilization cycles. Fertil Steril 80,350–355.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ and Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56,1129–1135.

Gougeon A (1984) Caracteres qualitatifs et quantitatifs de la population folliculaire dans l'ovaire humaine adulte. Contracept Fertil Sex 12(3),527–535.

Greb RR, Grieshaber K, Gromoll J, Sonntag B, Nieschlag E, Kiesel L and Simoni M (2005) A common single nucleotide polymorphism in exon 10 of the human follicle stimulating hormone receptor is a major determinant of length and hormonal dynamics of the menstrual cycle. J Clin Endocrinol Metab 90,4866–4872.

Grimes DA and Schulz KF (2005) Refining clinical diagnosis with likelihood ratios. Lancet 365,1500–1505.

Gulekli B, Bulbul Y, Onvural A, Yorukoglu K, Posaci C, Demir N and Erten O (1999) Accuracy of ovarian reserve tests. Hum Reprod 14,2822–2826.

Gurgan T, Urman B, Yarali H and Duran HE (1997) Follicle-stimulating hormone levels on cycle day 3 to predict ovarian response in women undergoing controlled ovarian hyperstimulation for in vitro fertilization using a flare-up protocol. Fertil Steril 68,483–487.

Hall JE, Welt CK and Cramer DW (1999) Inhibin A and inhibin B reflect ovarian function in assisted reproduction but are less useful at predicting outcome. Hum Reprod 14,409–415.

Hazout A, Bouchard P, Seifer DB, Aussage P, Junca AM and Cohen-Bacrie P (2004) Serum antimullerian hormone/mullerian-inhibiting substance appears to be a more discriminatory marker of assisted reproductive technology outcome than follicle-stimulating hormone, inhibin B, or estradiol. Fertil Steril 82,1323–1329.

Hendriks DJ, Broekmans FJ, Bancsi LF, de Jong FH, Looman CW and te Velde ER (2005a) Repeated clomiphene citrate challenge testing in the prediction of outcome in IVF: a comparison with basal markers for ovarian reserve. Hum Reprod 20,163–169.

Hendriks DJ, Broekmans FJ, Bancsi LF, Looman CW, de Jong FH and te Velde ER (2005b) Single and repeated GnRH agonist stimulation tests compared with basal markers of ovarian reserve in the prediction of outcome in IVF. J Assist Reprod Genet 22,65–73.

Hendriks DJ, Mol BW, Bancsi LF, te Velde ER and Broekmans FJ (2005c) Antral follicle count in the prediction of poor ovarian response and pregnancy after in vitro fertilization: a meta-analysis and comparison with basal follicle-stimulating hormone level. Fertil Steril 83,291–301.

Hendriks DJ, te Velde ER, Looman CW, Bancsi LF and Broekmans FJ (2005d). The role of poor response in the prediction of the cumulative ongoing pregnancy rate in in vitro fertilisation. Dynamic and basal ovarian reserve tests for outcome prediction in IVF: comparisons and meta-analyses. Academic Thesis, Utrecht, 162–179.

Honest H and Khan KS (2002) Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Serv Res 2,4.

Hsieh YY, Chang CC and Tsai HD (2001) Antral follicle counting in predicting the retrieved oocyte number after ovarian hyperstimulation. J Assist Reprod Genet 18,320–324.

Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL and te Velde ER (2004) Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. Hum Reprod 19,2019–2026.

Hunault CC, Laven JS, van Rooij IA, Eijkemans MJ te Velde ER and Habbema JD (2005) Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. Hum Reprod 20,1636–1641.

Huyser C, Fourie FL, Pentz J and Hurter P (1995) The predictive value of basal follicle stimulating and growth hormone levels as determined by immunofluorometry during assisted reproduction. J Assist Reprod Genet 12,244–251.

Irwig L, Macaskill P, Glasziou P and Fahey M (1995) Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 48,119–130.

Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC and Mosteller F (1994) Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120,667–676.

Jain T, Soules MR and Collins JA (2004) Comparison of basal follicle-stimulating hormone versus the clomiphene citrate challenge test for ovarian reserve screening. Fertil Steril 82,180–185.

Jarvela IY, Sladkevicius P, Kelly S, Ojha K, Campbell S and Nargund G (2003) Quantification of ovarian power Doppler signal with three-dimensional ultrasonography to predict response during in vitro fertilization. Obstet Gynecol 102,816–822.

Jinno M, Hoshiai T, Nakamura Y, Teruya K and Tsunoda T (2000) A novel method for assessing assisted female fertility: bioelectric impedance. J Clin Endocrinol Metab 85,471–474.

Kahraman S, Vicdan K, Isik AZ, Ozgun OD, Alaybeyoglu L, Polat G and Biberoglu K (1997) Clomiphene citrate challenge test in the assessment of ovarian reserve before controlled ovarian hyperstimulation for intracytoplasmic sperm injection. Eur J Obstet Gynecol Reprod Biol 73,177–182.

Khalifa E, Toner JP, Muasher SJ, Acosta AA (1992). Significance of basal follicle-stimulating hormone levels in women with one ovary in a program of in vitro fertilization. Fertil Steril 57,835–839.

Kim SH, Ku SY, Jee BC, Suh CS, Moon SY and Lee JY (2002) Clinical significance of transvaginal color Doppler ultrasonography of the ovarian artery as a predictor of ovarian response in controlled ovarian hyperstimulation for in vitro fertilization and embryo transfer. J Assist Reprod Genet 19,103–112.

Klinkert ER (2005) Clinical significance and management of poor response in IVF. Academic Thesis, Utrecht.

Klinkert ER, Broekmans FJ Looman CW and te Velde ER (2004) A poor response in the first in vitro fertilization cycle is not necessarily related to a poor prognosis in subsequent cycles. Fertil Steril 81,1247–1253.

Klinkert ER, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2005a) Expected poor responders on the basis of an antral follicle count do not benefit from a higher starting dose of gonadotrophins in IVF treatment: a randomized controlled trial. Hum Reprod 20,611–615.

Klinkert ER, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2005b) The antral follicle count is a better marker than basal follicle-stimulating hormone for the selection of older patients with acceptable pregnancy prospects after in vitro fertilization. Fertil Steril 83,811–814.

de Koning CH, Benjamins T, Harms P, Homburg R, van Montfrans JM, Gromoll J, Simoni M and Lambalk CB (2006) The distribution of FSH receptor isoforms is related to basal FSH levels in subfertile women with normal menstrual cycles. Hum Reprod 21,443–446.

van Kooij RJ, Looman CW, Habbema JD, Dorland M and te Velde ER (1996) Age-dependent decrease in embryo implantation rate after in vitro fertilization. Fertil Steril 66,769–775.

Kupesic S and Kurjak A (2002) Predictors of IVF outcome by three-dimensional ultrasound. Hum Reprod 17,950–955.

Kupesic S, Kurjak A, Bjelos D and Vujisic S (2003) Three-dimensional ultrasonographic ovarian measurements and in vitro fertilization outcome are related to age. Fertil Steril 79,190–197.

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB and Schoemaker J (2003) Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study. Hum Reprod 18,1422–1427.

Lambalk CB, de Koning CH, Flett A, van Kasteren Y, Gosden R and Homburg R (2004) Assessment of ovarian reserve. Ovarian biopsy is not a valid method for the prediction of ovarian reserve. Hum Reprod 19,1055–1059.

Lass A (2001) Assessment of ovarian reserve – is there a role for ovarian biopsy? Hum Reprod 16,1055–1057.

Lass A (2004) Assessment of ovarian reserve: is there still a role for ovarian biopsy in the light of new data? Hum Reprod 19,467–469.

Lass A, Silye R, Abrams DC, Krausz T, Hovatta O, Margara R and Winston RM (1997a) Follicular density in ovarian biopsy of infertile women: a novel method to assess ovarian reserve. Hum Reprod 12,1028–1031.

Lass A, Skull J, McVeigh E, Margara R and Winston RM (1997b) Measurement of ovarian volume by transvaginal sonography before ovulation induction with human menopausal gonadotrophin for in- vitro fertilization can predict poor response. Hum Reprod 12,294–297.

Lawson R, El Toukhy T, Kassab A, Taylor A, Braude P, Parsons J and Seed P (2003) Poor response to ovulation induction is a stronger predictor of early menopause than elevated basal FSH: a life table analysis. Hum Reprod 18,527–533.

Leridon H (1998) [30 years of contraception in France]. Contracept Fertil Sex 26,435–438.

Levi AJ, Raynault MF, Bergh PA, Drews MR, Miller BT and Scott RT (2001) Reproductive outcome in patients with diminished ovarian reserve. Fertil Steril 76,666–669.

Licciardi FL, Liu HC and Rosenwaks Z (1995) Day 3 estradiol serum concentrations as prognosticators of ovarian stimulation response and pregnancy outcome in patients undergoing in vitro fertilization. Fertil Steril 64,991–994.

Littenberg B and Moses LE (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 13,313–321.

Loumaye E, Billion JM, Mine JM, Psalti I, Pensis M and Thomas K (1990) Prediction of individual response to controlled ovarian hyperstimulation by means of a clomiphene citrate challenge test. Fertil Steril 53,295–301.

Martin JS, Nisker JA, Tummon IS, Daniel SA, Auckland JL and Feyles V (1996) Future in vitro fertilization pregnancy potential of women with variably elevated day 3 follicle-stimulating hormone levels. Fertil Steril 65,1238–1240.

Midgette AS, Stukel TA and Littenberg B (1993) A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making 13,253–257.

Mikkelsen AL, Andersson AM, Skakkebaek NE and Lindenberg S (2001) Basal concentrations of oestradiol may predict the outcome of in-vitro maturation in regularly menstruating women. Hum Reprod 16,862–867.

Mol BW, Dijkman B, Wertheim P, Lijmer J, van d V and Bossuyt PM (1997) The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. Fertil Steril 67,1031–1037.

Moses LE, Shapiro D and Littenberg B (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 12,1293–1316.

Muttukrishna S, Suharjono H, McGarrigle H and Sathanandan M (2004) Inhibin B and anti-Mullerian hormone: markers of ovarian response in IVF/ICSI patients? BJOG 111,1248–1253.

Muttukrishna S, McGarrigle H, Wakim R, Khadum I, Ranieri DM and Serhal P (2005) Antral follicle count, anti-mullerian hormone and inhibin B: predictors of ovarian response in assisted reproductive technology? BJOG 112,1384–1390.

Nahum R, Shifren JL, Chang Y, Leykin L, Isaacson K and Toth TL (2001) Antral follicle assessment as a tool for predicting outcome in IVF – is it a better predictor than age and FSH? J Assist Reprod Genet 18,151–155.

National Collaborating Center for Women's and Children's Health. (2004) *Fertility: Assessment and Treatment for People with Fertility Problems*. RCOG press, UK.

Ng EH, Tang OS and Ho PC (2000) The significance of the number of antral follicles prior to stimulation in predicting ovarian responses in an IVF programme. Hum Reprod 15,1937–1942.

Padilla SL, Bayati J and Garcia JE (1990) Prognostic value of the early serum estradiol response to leuprolide acetate in in vitro fertilization. Fertil Steril 53,288–294.

Penarrubia J, Balasch J, Fabregues F, Carmona F, Casamitjana R, Moreno V, Calafell JM and Vanrell JA (2000) Day 5 inhibin B serum concentrations as predictors of assisted reproductive technology outcome

in cycles stimulated with gonadotrophin-releasing hormone agonist-gonadotrophin treatment. Hum Reprod 15,1499–1504.

Penarrubia J, Fabregues F, Manau D, Creus M, Casals G, Casamitjana R, Carmona F, Vanrell JA and Balasch J (2005) Basal and stimulation day 5 anti-Mullerian hormone serum concentrations as predictors of ovarian response and pregnancy in assisted reproductive technology cycles stimulated with gonadotropin-releasing hormone agonist – gonadotropin treatment. Hum Reprod 20,915–922.

Perez MM, Gromoll J, Behre HM, Gassner C, Nieschlag E and Simoni M (2000) Ovarian response to follicle-stimulating hormone (FSH) stimulation depends on the FSH receptor genotype. J Clin Endocrinol Metab 85,3365–3369.

Phophong P, Ranieri DM, Khadum I, Meo F and Serhal P (2000) Basal 17beta-estradiol did not correlate with ovarian response and in vitro fertilization treatment outcome. Fertil Steril 74,1133–1136.

Popovic-Todorovic B, Loft A, Bredkjaeer HE, Bangsboll S, Nielsen IK and Andersen AN (2003a) A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. Hum Reprod 18,2275–2282.

Popovic-Todorovic B, Loft A, Lindhard A, Bangsboll S, Andersson AM and Andersen AN (2003b) A prospective study of predictive factors of ovarian response in 'standard' IVF/ICSI patients treated with recombinant FSH. A suggestion for a recombinant FSH dosage normogram. Hum Reprod 18,781–787.

Pruksananonda K, Boonkasemsanti W and Virutamasen P (1996) Basal follicle – stimulating hormone levels on day 3 of previous cycle are predictive of in vitro fertilization outcome. J Med Assoc Thai 79,365–369.

Qu J, Godin PA, Nisolle M and Donnez J (2000) Distribution and epidermal growth factor receptor expression of primordial follicles in human ovarian tissue before and after cryopreservation. Hum Reprod 15,302–310.

Ranieri DM, Quinn F, Makhlouf A, Khadum I, Ghutmi W, McGarrigle H, Davies M and Serhal P (1998) Simultaneous evaluation of basal follicle-stimulating hormone and 17 beta-estradiol response to gonadotropin-releasing hormone analogue stimulation: an improved predictor of ovarian reserve. Fertil Steril 70,227–233.

Ranieri DM, Phophong P, Khadum I, Meo F, Davis C and Serhal P (2001) Simultaneous evaluation of basal FSH and oestradiol response to GnRH analogue (F-G-test) allows effective drug regimen selection for IVF. Hum Reprod 16,673–675.

Roberts JE, Spandorfer S, Fasouliotis SJ, Kashyap S and Rosenwaks Z (2005) Taking a basal follicle-stimulating hormone history is essential before initiating in vitro fertilization. Fertil Steril 83,37–41.

Schild R L, Knobloch C, Dorn C, Fimmers R, van d V and Hansmann M (2001). The role of ovarian volume in an in vitro fertilization programme as assessed by 3D ultrasound. Arch Gynecol Obstet 265,67–72.

Schmidt KL, Ernst E, Byskov AG, Nyboe AA and Yding AC (2003) Survival of primordial follicles following prolonged transportation of ovarian tissue prior to cryopreservation. Hum Reprod 18,2654–2659.

Scott RT Jr and Hofmann GE (1995) Prognostic assessment of ovarian reserve [see comments]. Fertil Steril 63,1–11.

Scott RT, Toner JP, Muasher SJ, Oehninger S, Robinson S and Rosenwaks Z (1989) Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. Fertil Steril 51,651–654.

Seifer DB, Lambert Messerlian G, Hogan JW, Gardiner AC, Blazar AS and Berk CA (1997) Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome [see comments]. Fertil Steril 67,110–114.

Sharara FI and McClamrock HD (1999) The effect of aging on ovarian volume measurements in infertile women. Obstet Gynecol 94,57–60.

Sharara FI and McClamrock HD (2000) Antral follicle count and ovarian volume predict IVF outcome. Fertil Steril 74,S176.

Sharara FI and Scott RT (2004) Assessment of ovarian reserve. Is there still a role for ovarian biopsy? First do no harm! Hum Reprod 19,470–471.

Sharif K, Elgendy M, Lashen H and Afnan M (1998) Age and basal follicle stimulating hormone as predictors of in vitro fertilisation outcome. Br J Obstet Gynaecol 105,107–112.

Sharma V, Allgar V and Rajkhowa M (2002) Factors influencing the cumulative conception rate and discontinuation of in vitro fertilization treatment for infertility. Fertil Steril 78,40–46.

Smeenk JM, Stolwijk AM, Kremer JA and Braat DD (2000) External validation of the templeton model for predicting success after IVF. Hum Reprod 15,1065–1068.

Smotrich DB, Widra EA, Gindoff PR, Levy MJ, Hall JL and Stillman RJ (1995) Prognostic value of day 3 estradiol on in vitro fertilization outcome. Fertil Steril 64,1136–1140.

Snick HK, Snick TS, Evers JL and Collins JA (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. Hum Reprod 12,1582–1588.

Spira A (1988) The decline of fecundity with age. Raturitas 10 (Suppl),15–22.

Stolwijk AM, Zielhuis GA, Hamilton CJ, Straatman H, Hollanders JM, Goverde HJ, van Dop PA and Verbeek AL (1996) Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. Hum Reprod 11,2298–2303.

Stolwijk AM, Straatman H, Zielhuis GA, Jansen CA, Braat DD, van Dop PA and Verbeek AL (1998) External validation of prognostic models for ongoing pregnancy after in-vitro fertilization. Hum Reprod 13,3542–3549.

Syrop CH, Willhoite A and Van Voorhis BJ (1995) Ovarian volume: a novel outcome predictor for assisted reproduction. Fertil Steril 64,1167–1171.

Tanbo T, Dale PO, Abyholm T and Stokke KT (1989) Follicle-stimulating hormone as a prognostic indicator in clomiphene citrate/human menopausal gonadotrophin-stimulated cycles for in-vitro fertilization. Hum Reprod 4,647–650.

Tanbo T, Abyholm T, Bjoro T and Dale PO (1990) Ovarian stimulation in previous failures from in-vitro fertilization: distinction of two Groups of poor responders. Hum Reprod 5,811–815.

Tanbo T, Dale PO, Lunde O, Norman N and Abyholm T (1992) Prediction of response to controlled ovarian hyperstimulation: a comparison of basal and clomiphene citrate-stimulated follicle- stimulating hormone levels. Fertil Steril 57,819–824.

Tarlatzis BC, Zepiridis L, Grimbizis G and Bontis J (2003) Clinical management of low ovarian response to stimulation for IVF: a systematic review. Hum Reprod Update 9,61–76.

Templeton A, Morris JK and Parslow W (1996) Factors that affect outcome of in-vitro fertilisation treatment [see comments]. Lancet 348,1402–1406.

Te Velde ER and Pearson PL (2002) The variability of female reproductive aging. Hum Reprod Update 8,141–154.

Toner JP, Philput CB, Jones GS and Muasher SJ (1991) Basal follicle-stimulating hormone level is a better predictor of in vitro fertilization performance than age. Fertil Steril 55,784–791.

Ulug U, Ben Shlomo I, Turan E, Erden HF, Akman MA and Bahceci M (2003) Conception rates following assisted reproduction in poor responder patients: a retrospective study in 300 consecutive cycles. Reprod Biomed Online 6,439–443.

Van der Stege JG and van der Linden PJ (2001) Useful predictors of ovarian stimulation response in women undergoing in vitro fertilization. Gynecol Obstet Invest 52,43–46.

Van Rooij IAJ, Broekmans FJ, te Velde ER, Fauser BCJM, Bancsi LFJMM, de Jong FH and Themmen APN (2002) Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod 17,101–107.

Van Rooij IA, Broekmans FJ, Hunault CC, Scheffer GJ, Eijkemans MJ, de Jong FH, Themmen AP and te Velde ER (2006) The use of ovarian reserve tests for the prediction of ongoing pregnancy in couples with unexplained female subfertility. Reprod Biomed Online 12,182–190.

Vazquez ME, Verez JR, Stern JJ, Gutierrez NA and Asch RH (1998) Elevated basal estradiol levels have no negative prognosis in young women undergoing ART cycles. Gynecol Endocrinol 12,155–159.

Ventura SJ, Mosher WD, Curtin SC, Abma JC and Henshaw S (2001) Trends in pregnancy rates for the United States 1976–97: an update. Natl Vital Stat Rep 49,1–9.

Webber LJ, Stubbs S, Stark J, Trew GH, Margara R, Hardy K and Franks S (2003) Formation and early development of follicles in the polycystic ovary. Lancet 362,1017–1021.

Weinstein M, Wood AJ and Chang MC (1993) Age patterns in fecundability. In Gray R, Leridon H and Spira A (eds) *Biomedical and Demographic Determinants of Reproduction*. Clarendon Press, Oxford, pp. 209–220.

Winslow KL, Toner JP, Brzyski RG, Oehninger SC, Acosta AA and Muasher SJ (1991) The gonadotropin-releasing hormone agonist stimulation test – a sensitive predictor of performance in the flare-up in vitro fertilization cycle. Fertil Steril 56,711–717.

Wood JW (1989) Fecundity and natural fertility in humans. Oxf Rev Reprod Biol 11,61–109.

Yanushpolsky EH, Hurwitz S, Tikh E and Racowsky C (2003) Predictive usefulness of cycle day 10 follicle-stimulating hormone level in a clomiphene citrate challenge test for in vitro fertilization outcome in women younger than 40 years of age. Fertil Steril 80,111–115.

Yong PY, Baird DT, Thong KJ, McNeilly AS and Anderson RA (2003) Prospective analysis of the relationships between the ovarian follicle cohort and basal FSH concentration, the inhibin response to exogenous FSH and ovarian follicle number at different stages of the normal menstrual cycle and after pituitary down-regulation. Hum Reprod 18,35–44.

Zaidi J, Barber J, Kyei Mensah A, Bekir J, Campbell S and Tan SL (1996). Relationship of ovarian stromal blood flow at the baseline ultrasound scan to subsequent follicular response in an in vitro fertilization program. Obstet Gynecol 88,779–784.

Chapter 8

General discussion, conclusions and recommendations for future research

The primary aim of this thesis was to identify a test or combination of tests, which can predict the cohort size of small antral follicles in the early follicular phase. The size of this cohort should reflect the number of developing mature follicles when stimulated with a superphysiological FSH dosage, for example during ovarian hyperstimulation in an IVF cycle.
The further aim of the study was to find a minimally invasive, reliable ovarian reserve test which can give a prediction for a poor, hyper or adequate response after ovarian hyperstimulation and a prognosis for pregnancy.
The relevance of research in this area lies within the fact that assisted reproduction techniques (ART) are complex and expensive and have strict indications. Informing infertile couples about their chances of pregnancy, naturally or by means of ART should have a high priority. Despite extensive research there is still no adequate test for the prediction of ovarian reserve and pregnancy. With the postponement of childbearing, the age related fertility decline has been shown to play an important role in the increase in infertility among couples who are trying to conceive. In IVF treatment this age effect has been shown in many accumulated data. Due to the variation of female fertility within a certain age category the need was felt for tests that better identified cases with a state of ovarian reserve that is clearly too low for their age.
Since a gold standard for ovarian reserve status in the sense of quantity and quality is lacking, the occurrence of poor ovarian response to maximal stimulation and the occurrence of pregnancy in IVF are used as parameters to assess the accuracy of the test. The ideal ovarian reserve test should identify a substantial percentage of IVF indicated cases, which have a practically zero chance of becoming pregnant in a series of treatment cycles due to the adverse effects of diminished ovarian reserve. Those cases can be prevented from entering the program, as they will engender very high costs for only minimal results. If not too expensive and not too demanding for the patient, such a test would be readily embraced by physicians, patients, health politicians and insurance companies.
In this section the research questions as formulated in *chapter 1* will be answered and discussed. First a concise answer will be given. After each answer, an extended motivation and discussion is provided in which we will also stress the issue of clinical value.

Question:	**Which ovarian reserve test or a certain combination can predict the cohort size of small antral follicles in the early follicular phase?**
Answer:	**The model with the variables: total antral follicle count, Inhibin B-increment in the EFORT and the total basal volume of the ovaries best fulfilled the credentials as the ideal model. The regression line of the total antral follicle count, Inhibin B-increment and total basal volume on the number of follicles was drawn by the regression equation: Y = -3.161 + 0.805 x AFC (0.258-1.352) + 0.034 x Inh. B-incr.(0.007-0.601) + 0.511 BOV (0.480-0.974)(r=0.848,p<0.001).**

Reproductive aging is thought to be dictated by a gradual decrease in both the quantity and the quality of the oocytes and follicles held within the ovaries (te Velde *et al.,* 1998, te Velde and Pearson, 2002). With regard to quantity, histological studies have shown that at birth a few million primordial follicles are present from which at the onset of puberty only some 400.000 are left (Block, 1952, Block, 1953, Faddy *et al.,* 1992, Gougeon, 1998). The wasting of follicles continuously throughout reproductive life, reaching a critical number of a few thousand at a mean age of 45 when menstrual cycles become irregular, and falling to clearly below a thousand follicles at the time menstrual cycles cease, the event know as menopause (Gosden and Faddy, 1994, Richardson *et al.*, 1990, Richardson *et al.*, 1987). In analogy to these histological changes, Scheffer *et al.* (1999) demonstrated that the number of primordial follicles in the ovary, as published by Faddy and Gosden. (1996) correlated well with the number of growing follicles, counted by transvaginal sonography in the early follicular phase. So the decreasing size of the antral follicle cohort with age is a reflection of the decreasing primordial follicle pool. We used this principle to measure ovarian reserve, defined as the total number of follicles which can be stimulated under maximal ovarian stimulation with FSH. A number of the so-called ovarian reserve tests are supposed to indirectly reflect the size of the cohort of small antral follicles (2-10 mm in diameter) in the ovary. This decrease in follicle number is exemplified by the increased risk of producing a poor response in ovarian hyperstimulation in IVF patients at older age (Goverde *et al.,* 2005, Akande *et al.,* 2004, Beckers *et al.,* 2002).

The first noticeable clinical sign for the advancement in the reproductive ageing process is a shortening of the menstrual cycle length by some 2-3 days while regularity remains unaffected (Treloar *et al.,* 1967). This is predominantly the result of a shortening of the follicular phase. It seems that growing speed of the dominant follicle is not different in older compared to younger women (Lambalk *et al.,* 1998, Klein *et al.,* 2002, van Zonneveld *et al.,* 2003), but may well be advanced and may have already started in the luteal phase of the preceding cycle, with a consequent early production of estradiol (E2). This premature E2 elevation signifies early recruitment and is a common perimenopausal pattern. Elevated E2 levels exert a negative feedback on the hypothalamic-pituitary axis, reducing FSH secretion. In *chapter 2* the bE2 was investigated to ascertain its ability to predict the FSH-sensitive follicle cohort. The bE2 was not found to be capable of predicting the number of FSH-sensitive follicles growing in a subsequent IVF cycle.

A rise in early follicular levels of follicle stimulating hormone (FSH) is considered as an endocrine sign of the aging ovary (Fitzgerald *et al.,* 1994, Klein *et al.,* 1993, van Zonneveld, 2001). It is caused by a reduction in the negative feedback on FSH secretion from the pituitary and results in an increased pituitary responsiveness to GnRH (de Koning *et al.,* 2000). This altered responsiveness is due to a decrease in levels of ovarian feedback hormones such as the inhibins and putatively by a decline of the granulosa cell gonadotropin surge inhibiting factor(GnSIF) (de Koning *et al.*, 2000). Elevated levels of FSH are consistent with already advanced stages of the ovarian ageing process (Levi *et al.,* 2001). In our study, the bFSH was not found to be capable of predicting the number of FSH-sensitive follicles growing in a subsequent IVF cycle (*chapter 2*).

Recently, anti-Müllerian hormone (AMH) produced by preantral and antral follicles up to a diameter of 6 mm has been shown to be a marker for the size of the residual antral follicle pool and its decline follows the ageing process in a more gradual fashion (van Rooij *et al.,* 2002, van Rooij *et al.,* 2004, van Rooij *et al.,* 2005, Fanchin *et al.,* 2003, Seifer *al.,* 2002).

In *chapter 6* the AMH was investigated to ascertain its ability to predict the FSH-sensitive follicle cohort. But AMH was not found to be a better test compared to the EFORT, AFC or BOV, for the prediction of the number of FSH-sensitive follicles growing in a subsequent IVF cycle.
In theory, the direct products of granulosa cells may best reflect ovarian secretory reserve and follicle number. Seifer *et al*. (1996a) reported that women undergoing IVF with high bFSH have preovulatory follicles with fewer luteinized granulosa cells and an increase in the percentage of cells undergoing apoptosis as compared to women with a low bFSH. With increased understanding of the control of synthesis and secretion of the inhibins and their potential endocrine role in the regulation of FSH in the human, attention has turned to the possibility that this family of peptides may provide a more direct index of ovarian reserve and improve predictions of IVF outcome (Seifer *et al.*, 1996b, Danforth *et al.*, 1998, Hall *et al.*, 1999, Seifer *et al.*, 1997). The decreased inhibin B levels in women of advanced reproductive age may be the result of a smaller cohort in the ovaries of older women, caused by fewer primordial to early antral follicles proceeding to the recruitment stage. Basal Inhibin B was not capable of predicting the FSH-sensitive follicle cohort after IVF stimulation (*chapter 2*).
We used this knowledge in the EFORT, in which the ability of the ovaries to respond to a fixed dose of exogenously administered FSH (300 IU FSH) was demonstrated on cyle day 3 in 24 hours. FSH induces aromatase activity. The aromatase activity results in increased follicular concentrations of estradiol since the aromatase substrate, androstenedione is abundantly available. Thereby the estradiol increase becomes a parameter for the size of a growing cohort. As has been suggested, granulosa cells of small antral follicles under the influence of FSH also produce Inhibin B. In *chapter 2* we showed that the Inhibin B-increment and E2-increment in the EFORT are the good predictors of the total number of follicles obtained after maximal ovarian hyperstimulation in an IVF-treatment i.e. cohort size. age, bFSH, bE2, bInhibin B and the outcome of the CCCT (bFSH + sFSH) in this respect each, and in combination, showed a much lower performance.
In conjunction, a gradual decrease with advancing age in the number of sonographically detectable antral follicles has been shown in many studies (Scheffer *et al.,* 2003, Gougeon, 1998). In recent years several papers have been published concerning the relation between the antral follicle count (AFC, defined as the total number of antral follicles, sized 2-5 or 2-10 mm, present in both ovaries) and the ovarian response in IVF (Bancsi *et al.*, 2002, Chang *et al.*, 1998), as well as the occurrence of the menopausal transition (van Rooij *et al.,* 2004), indicating that this parameter relates strongly to the quantitative aspects of ovarian reserve. In *chapter 5* the AFC and total basal ovarian volume were investigated to ascertain its ability to predict the FSH-sensitive follicle cohort. They both were found to be capable of predicting the number of FSH sensitive follicles growing in a subsequent IVF cycle. However, the correlation between ovarian volume and ovarian response is less significant than the correlation of AFC and ovarian response. The volume of the ovaries is an indirect indicator of the activity of the ovaries.
In *chapter 2, 5 and 6* we tried to build a model which represents the actual functional status of the ovaries. The model with the variables: total antral follicle count, Inhibin B-increment in the EFORT and the total basal volume of the ovaries best fulfilled the credentials as the ideal model.

The regression line of the total antral follicle count, Inhibin B-increment and total basal volume on the number of follicles was drawn by the regression equation: Y = -3.161 + 0.805 x AFC (0.258-1.352) + 0.034 x Inh. B-incr.(0.007-0.601) + 0.511 BOV (0.480-0.974)(r=0.848,p<0.001).
Inter-observer and intra-observer variation in AFC and ovarian volume measurements has been shown to be quite low (5-7%) (Bancsi *et al.*, 2004, Scheffer *et al.,* 2002). *In chapter 4* we showed that the intercycle variability of the Inhibin-B increment and the estradiol increment in the EFORT is stable in consecutive cycles, so this indicates that all variables in the model are highly reproducible and therefore this model is an excellent test to predict ovarian reserve in a quantitative way.
The clinical value of such a model is debatable. For the patient and physicians the unifying goals are traditionally to find out how a patient will respond to stimulation and what are their chances of pregnancy. For clinical relevance, test characteristics like sensitivity, specificity and predictive value must be calculated to identify dimished ovarian response, hyper ovarian response and the chances on pregnancy, which may potentially lead to adapting the stimulation protocol, e.g. by applying more or less aggressive stimulation with higher or lower doses of gonadotropins. Therefore we formulated the next question.

Question: **Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of poor and hyper ovarian response in an IVF population?**

Answer: **AFC is the best test for the prediction of poor and hyper responders. The predictive value of AMH for poor response is comparable with that of AFC, but unfortunately AMH is not a good test for the prediction of hyper responders. However AMH is probably more easily applicable in general practice, because it can be measured anywhere in the cycle. However, validation for this way of application of AMH as test for ovarian response needs to be validated.**

From the *systematic and meta-analytic review* and *chapter 3, 5 and 6*, presented in this thesis it can be concluded that the accuracy of the AFC for predicting poor and hyper response in regularly cycling women is adequate. Added to the false positive rate of ~5% the test will not be suitable as diagnostic test to exclude patients on the basis of the presumed diagnosis of advanced ovarian ageing. It may well be used as a screening test for possible poor responders and for directing further diagnostic steps like a first IVF attempt, where the ovarian response to hyperstimulation will provide additional information (Hendriks *et al.*, 2005b). The accuracy of AFC for the occurrence of poor ovarian response to hyperstimulation appeared to be good. This is understandable as the response to stimulation represents, the size of the cohort of FSH sensitive follicles continuously present in the ovaries and directly related to the size of the primordial follicle pool (Gougeon, 1984). The predictive value of AMH for poor response is comparable with that of AFC, but unfortunately AMH is not a good test for the prediction of hyper responders. However AMH is probably more easily applicable in general practice, because it can be measured anywhere in the cycle. However, validation for this way of application of AMH as test for ovarian response needs to be validated.

Poor ovarian response has been associated with a reduced chance of pregnancy in the actual treatment cycle as well as in subsequent cycles and as such may well be indicative of ovarian reserve status in both the quantitative and qualitative sense (Klinkert *et al.*, 2004, Klinkert *et al.,* 2005b, Ulug *et al.,* 2003). Accurate prediction of poor response could therefore have clinical value if the pregnancy prospects are so unfavorable that a predicted poor responder would be denied treatment. Accuracy in response prediction, however, will only be high if the false positives are prevented by using extreme cut off levels, implying that only minor percentages of abnormal tests will be found and many future poor responders will pass unrecognized. At the same time it is necessary to know whether the predicted poor responder indeed has very low prospects for success in subsequent cycles. As much of this is unknown at the present time, the use of any ovarian reserve test for poor response prediction can not be supported, not even if it would be used for adapting the treatment schedule in anticipated poor responders, as no altered treatment schedule has consistently shown to be effective in women with a severely reduced size of the follicle cohort (Tarlatzis *et al.*, 2003, Klinkert *et al.*, 2005a Klinkert *et al.*, 2005)

Question:	**Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of pregnancy in an IVF population?**
Answer:	**Non of the ovarian reserve tests with an acceptable cutoff level served as a good predictor for pregnancy in an IVF population.**

From the *systematic and meta-analytic review* and *chapter 6*, presented in this thesis it can be concluded that the ovarian reserve tests known to date have very modest predictive properties for the prediction of the occurrence of pregnancy and are therefore not suitable for relevant clinical use. Only if a high cutoff is used, in order to prevent couples from wrongly being refused IVF, a very small minority of IVF indicated cases (~3%) is identified as having unfavorable prospects in an IVF treatment cycle (pregnancy rate for that cycle $\leq$ 5%). It should be noted that the use of pregnancy as outcome parameter for the assessment of ovarian reserve status may be insufficient if only one exposure cycle is taken into account. As such, the possibility of misjudgment on the basis of currently known ovarian reserve tests is hard to rule out. Also, it has been shown sufficiently that women with signs of a diminished ovarian reserve may still become pregnant (van Montfrans *et al*., 2004, Van Rooij *et al*., 2004). This implies that the use of the test as a method to deny treatment to assumed ovarian aged women should be declined and, as a consequence, the test should not be applied on a regular basis or only used for counseling.

Conclusions

For the prediction of the cohort size of small antral follicles in the early follicular phase in a quantitative way, we could built a nice model with 3 ovarian reserve tests, but this is less usable in clinical practice. The ideal ovarian reserve test for physicians should identify a substantial percentage of IVF indicated cases, who have a practically zero chance of becoming pregnant in an IVF programme due to the adverse effects of dimished ovarian reserve. Those

cases can be refrained from entering the program, as they will raise very high costs for only minimal results. If not too expensive and not too demanding for the patient, such a test would be readily embraced by physicians, patients, health politicians and insurance companies.
From the prospective study (*chapter 2-6*) in this thesis and the systematic reviews (*chapter 7*) presented in this thesis it can be concluded that the ovarian reserve tests known to date have only minimal predictive properties and are therefor far from suitable for relevant clinical use. This implies that the use of the ovarian reserve test as a method to deny treatment to assumed ovarian aged women should be declined and, as a consequence the test should not be applied on a regular basis. Also for patients, the value of testing for ovarian reserve seems not useful for current IVF programmes. This is based on a decision model, where current test accuracy and preference inventory among patients and physicians were used (Mol *et al*., 2006).
As poor ovarian response will provide some information on ovarian reserve status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy. Once a poor response is obtained, the question arises whether this finding is based on depleted ovaries or other causes, like underdosing for instance based on the presence of certain FSH receptor polymorphisms (Perez *et al.,* 2000, de Koning *et al.,* 2005, Behre *et al.,* 2005, Greb *et al.,* 2005). A repeat cycle with adequate, maximal stimulation or a post hoc performed ovarian reserve test (basal FSH or AFC (Hendriks *et al.*, 2005a)) may correctly classify the poor responder patient as ovarian aged and allow properly based refraining from further treatment (Klinkert *et al.*, 2004).
It should be remembered that the purpose of any ovarian reserve test is the identification of women with poor ovarian reserve for her age. This implies that chronological age always is the first step in ovarian reserve assessment. In young women ovarian reserve tests may help to classify poor responders and direct management in these cases by estimating the size of the FSH sensitive cohort. In older women ovarian reserve tests may help to identify those cases that, in spite of their age, still may have acceptable chances of becoming pregnant through IVF as the quantity of response to stimulation is anticipated to be normal or even high (Klinkert *et al.,* 2005b).

Future perspectives
The occurrence of menopause in the female represents an almost exhausted primordial and antral follicle pool (Faddy *et al.*, 1992). At the same time it is an event that can rather easily be recognized in individuals. The mean age at menopause is 51 years with variation ranging from ~40 – ~60 years (Treloar, 1981), but possibly even into younger age ranges (Broekmans *et al.*, 2004). This range of variation seems also to be true for the occurrence of the cycle irregularities which are so typical for the menopausal transition (den Tonkelaar *et al.,* 1998). Between the occurrence of cycle irregularity and the occurrence of menopause a fixed temporal relationship is believed to be present (te Velde and Pearson, 2002). As the age at which a woman becomes naturally sterile shows the same degree of variation as observed for menopause and the occurrence of cycle irregularity (Broekmans *et al.*, 2004), it is generally assumed that this event, which obviously is not easily recognizable, carries an individually fixed temporal relationship with cycle irregularity and menopause with a presumed interval of about 5 and 10 years, respectively. Whether the same hypothetical relationship is true for the age at which women start to become subfertile (assumed mean age 31 years) remains to be elucidated. Recently, from population studies first evidence for such a relationship has been obtained (Eijkemans *et al.,* 2005).

From research on factors that contribute to the variation in the cascade of reproductive events it has been shown that various lifestyle and environmental factors contribute only slightly to this variation (Gold *et al.,* 2001, McKinlay *et al.,* 1985, Thomas *et al.,* 2001, Van Noord *et al.,* 1997). In contrast, inheritance has been shown to be the major factor that determines the large variation in the occurrence of reproductive events (Do *et al.,* 2000, Treloar *et al.,* 2000, de Bruin *et al.,* 2001). Given the concept of the high heritability of age at menopause, genes that determine the number of the follicles that develop in the ovaries during fetal life and the rate of loss of these follicles by atresia thereafter are believed to be the same genes that may be responsible for the variation in the age at menopause and the occurrence of subfertility and sterility (van Asselt *et al.,* 2004a, van Asselt *et al.,* 2004b). It is assumed that the heritability in reproductive events follows the pattern of multifactorial inheritance where genes interact with environmental factors (van Asselt *et al.*, 2003). It is to be expected that from approaches like sibpair and linkage analysis candidate genes will be discovered that further direct the research in this field (van Asselt *et al.*, 2003).
So future perspectives in this research field can be found in studies where success rates in cumulative treatment cycles or in units of time (one year treatment period) are analyzed as to the question whether any test will correctly identify those couples that do not become pregnant in such series of exposures. Novel tests that most accurately estimate the age at which menopause is expected to take place in an individual woman may facilitate the estimation of the remaining reproductive potential at a certain age. Such a test will probably be based on family history taking (age at menopause of mother) or will comprise testing for genetic markers, that may be discovered from large scale, population genetic screening.

References

Akande VA, Keay SD, Hunt LP et al (2004) The practical implications of a raised serum FSH and age on the risk of IVF treatment cancellation due to a poor ovarian response. J Assist Reprod Genet 21, 257-262.

Bancsi LFJMM, Broekmans FJM, Eijkemans MJC et al (2002) Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 77, 328-336.

Bancsi LF, Broekmans FJ, Looman CW et al (2004a) Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. Fertil Steril, 81, 35-41.

Beckers NG, Macklon NS, Eijkemans MJ et al (2002) Women with regular menstrual cycles and a poor response to ovarian hyperstimulation for in vitro fertilization exhibit follicular phase characteristics suggestive of ovarian aging. Fertil Steril 78, 291-297.

Behre HM, GrebRR, Mempel A et al (2005) Significance of a common single nucleotide polymorphism in exon 10 of the follicle-stimulating hormone (FSH) receptor gene for the ovarian response to FSH: a pharmacogenetic approach to controlled ovarian hyperstimulation. Pharmacogenet Genomics 15, 451-456.

Block E (1952) Quantitative morphological investigations of the follicular system in women. Variations at different ages. Acta Anat 14, 108-23.

Block E (1953). A quantitative morphological investigation of the follicular system in newborn female infants. Acta Anat 17, 201-206.

Broekmans FJ, Faddy MJ, Scheffer G et al (2004b) Antral follicle counts are related to age at natural fertility loss and age at menopause. Menopause 11, 607-614.

Chang MY, Chiang CH, Hsieh TT, Soong YK, and Hsu KH(1998b) Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil Steril 69, 505-510.

Danforth DR, Arbogast LK, Mroueh J, Kim MH, Kennard EA, Seifer DB et al (1998). Dimeric inhibin: a direct marker of ovarian aging. Fertil Steril 70,119-23.

de Bruin JP, Bovenhuis H, Van Noord PA et al (2001) The role of genetic factors in age at natural menopause. Hum Reprod 16, 2014-2018.

de Koning CH, Popp-Snijders C, Schoemaker J et al (2000b) Elevated FSH concentrations in imminent ovarian failure are associated with higher FSH and LH pulse amplitude and response to GnRH. Hum Reprod 15, 1452-1456.

de Koning CH, Benjamins T, Harms P et al (2005) The distribution of FSH receptor isoforms is related to basal FSH levels in subfertile women with normal menstrual cycles. Hum.Reprod 21, 443-6.

den Tonkelaar I, te Velde ER, and Looman CW (1998) Menstrual cycle length preceding menopause in relation to age at menopause. Am J Human Biol 29, 115-123.

Do KA, Broom BM, Kuhnert P et al (2000) Genetic analysis of the age at menopause by using estimating equations and Bayesian random effects models. Stat Med 19, 1217-1235.

Eijkemans MJ, Habbema JDF, and te Velde ER. Age at last childbirth and fertility at young age. In: Fertility in Populations and in Patients; M.J. Eijkemans; Academic Thesis, Rotterdam 23-34, 2005.

Faddy MJ, Gosden RG, Gougeon A et al (1992) Accelerated disappearance of ovarian follicles in mid-life: implications for forecasting menopause. Hum Reprod 7, 1342-1346.

Faddy MJ, Gosden RG (1996) A model conforming the decline in follicle numbers to the age of menopause in women. Hum Reprod 11, 1484-6.

Fanchin R, Schonauer LM, Righini C, frydman N, Frydman R, Taieb J (2003) Serum anti-Mullerian hormone dynamics during controlled ovarian hyperstimulation. Hum Reprod 18, 328-32.

Fitzgerald CT, Seif M, Killick S et al (1994) Age related changes in the female reproductive cycle [published erratum appears in Br J Obstet Gynaecol 1994, 101, 360]. Br J Obstet Gynaecol 101, 229-233.

Gold EB, Bromberger J, Crawford S et al (2001) Factors associated with age at natural menopause in a multiethnic sample of midlife women. Am J Epidemiol 153, 865-874.

Gosden RG and Faddy MJ (1994) Ovarian aging, follicular depletion, and steroidogenesis. Exp Gerontol 29, 265-274.

Gougeon A (1984). Caracteres qualitatifs et quantitatifs de la population folliculaire dans l'ovaire humaine adulte. Contracept Fertil Sex 12, 527-535.

Gougeon A. (1998) Ovarian follicular growth in humans: ovarian ageing and population of growing follicles. Am J Human Biol 30, 137-142.

Goverde AJ, McDonnell J, Schats R et al (2005) Ovarian response to standard gonadotrophin stimulation for IVF is decreased not only in older but also in younger women in couples with idiopathic and male subfertility. Hum Reprod 20, 1573-1577.

Greb RR, Grieshaber K, Gromoll J et al (2005) A common single nucleotide polymorphism in exon 10 of the human follicle stimulating hormone receptor is a major determinant of length and hormonal dynamics of the menstrual cycle. J Clin Endocrinol Metab 90, 4866-4872.

Hall JE, Welt CK, Cramer DW (1999). Inhibin A and inhibin B reflect ovarian function in assisted reproduction but are less useful at predicting outcome. Hum Reprod 4, 409-15.

Hendriks DJ, Mol BW, Bancsi LF et al (2005a) Antral follicle count in the prediction of poor ovarian response and pregnancy after in vitro fertilization: a meta-analysis and comparison with basal follicle-stimulating hormone level. Fertil Steril 83, 291-301.

Hendriks DJ, te Velde ER, Looman CW, Bancsi LF and Broekmans FJ (2005b). The role of poor response in the prediction of the cumulative ongoing pregnancy rate in in vitro fertilisation. Dynamic and basal ovarian reserve tests for outcome prediction in IVF: comparisons and meta-analyses. Academic Thesis, Utrecht, 162–179.

Klein NA, Pergola GM, Rao Tekmal R et al (1993) Enhanced expression of resident leukocyte interferon gamma mRNA in endometriosis. Am J Reprod Immunol, 30, 74-81.

Klein NA, Harper AJ, Houmard BS et al (2002) Is the short follicular phase in older women secondary to advanced or accelerated dominant follicle development? J Clin Endocrinol Metab 87, 5746-5750.

Klinkert ER. Clinical significance and management of poor response in IVF. Academic Thesis, Utrecht 2005.

Klinkert ER, Broekmans FJ, Looman CW et al (2005a) Expected poor responders on the basis of an antral follicle count do not benefit from a higher starting dose of gonadotrophins in IVF treatment: a randomized controlled trial. Hum Reprod 20, 611-615.

Klinkert ER, Broekmans FJ, Looman CW et al (2005b) The antral follicle count is a better marker than basal follicle-stimulating hormone for the selection of older patients with acceptable pregnancy prospects after in vitro fertilization. Fertil Steril 83, 811-814.

Klinkert ER, Broekmans FJ, Looman CW et al (2004) A poor response in the first in vitro fertilization cycle is not necessarily related to a poor prognosis in subsequent cycles. Fertil Steril 81, 1247-1253.

Lambalk CB and de Koning CH (1998) Interpretation of elevated FSH in the regular menstrual cycle. Am J Human Biol 30, 215-220.

Levi AJ, Raynault MF, Bergh PA et al (2001) Reproductive outcome in patients with diminished ovarian reserve. Fertil Steril 76, 666-669.

McKinlay SM, Bifano NL, and McKinlay JB (1985) Smoking and age at menopause in women. Ann Intern Med 103, 350-356.

Mol BW, Verhagen TE, Hendriks DJ, Collins JA, Coomarasamy A, Opmeer BC, Broekmans FJ (2006)Value of ovarian reserve testing before IVF: a clinical decision analysis. Hum Reprod 21, 1816-1823.

Perez MM, Gromoll J, Behre HM et al (2000) Ovarian response to follicle-stimulating hormone (FSH) stimulation depends on the FSH receptor genotype. J Clin Endocrinol Metab 85, 3365-3369.

Richardson SJ and Nelson JF (1990) Follicular depletion during the menopausal transition. Ann NY Acad Sci 592, 13-20.

Richardson SJ, Senikas V, and Nelson JF (1987) Follicular depletion during the menopausal transition: evidence for accelerated loss and ultimate exhaustion. J Clin Endocrinol Metab 65, 1231-1237.

Scheffer GJ, Broekmans FJ, Dorland M et al (1999) Antral follicle counts by transvaginal ultrasonography are related to age in women with proven natural fertility. Fertil Steril 72, 845-851.

Scheffer GJ, Broekmans FJ, Bancsi LF et al (2002) Quantitative transvaginal two- and three-dimensional sonography of the ovaries: reproducibility of antral follicle counts. Ultrasound Obstet Gynecol 20, 270-275.

Scheffer GJ, Broekmans FJ, Looman CW, Blankenstein MA, Fauser BC, te Jong FH and te VeldeER (2003) The number of antral follicles measured by transvaginal ultrasound (TVS) is the best predictor of reproductive age in a group of normal fertile women. Hum Reprod 18, 700-6.

Seifer DB, Gardiner AC, Ferreira KA, Peluso JJ (1996a) Apoptosis as a function of ovarian reserve in women undergoing in vitro fertilization. Fertil Steril 66, 593-8

Seifer DB, Gardiner AC, Lambert-Messerlian G, Schneyer AL (1996b) Differential secretion of dimeric inhibin in cultured luteinized granulosa cells as a function of ovarian reserve. J Clin Endocrinol Metab 81, 736-9.

Seifer DB, Lambert-Messerlian G, Hogan JW (1997) Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome. Fertil Steril 67, 110-4.

Seifer DB, Mac Laughlin DT, Christian BP, Feng B and Shelden RM (2002) Early follicular serum mullerian-inhibiting substance levels are associated with ovarian response during assisted reproductive technology cycles. Fertil Steril 77, 468-71.

Tarlatzis BC, Zepiridis L, Grimbizis G et al (2003) Clinical management of low ovarian response to stimulation for IVF: a systematic review. Hum Reprod Update 9, 61-76.

Te Velde ER, Dorland M, Broekmans FJ (1998). Age at menopause as a marker of reproductive ageing. Maturitas 30, 119-25.

Te Velde ER and Pearson PL (2002) The variability of female reproductive ageing. Hum Reprod Update 8, 141-154.

Thomas F, Renaud F, Benefice E et al (2001) International variability of ages at menarche and menopause: patterns and main determinants. Hum Biol 73, 271-290.

Treloar AE, Boynton RE, Behn BG et al (1967) Variation of the human menstrual cycle through reproductive life. Int J Fertil 12, 77-126.

Treloar AE (1981) Menstrual cyclicity and the pre-menopause. Am JHuman Biol 3, 249-264.

Treloar SA, Sadrzadeh S, Do KA et al (2000) Birth weight and age at menopause in Australian female twin pairs: exploration of the fetal origin hypothesis. Hum Reprod 15, 55-59.

Ulug U, Ben Shlomo I, Turan E et al (2003) Conception rates following assisted reproduction in poor responder patients: a retrospective study in 300 consecutive cycles. Reprod Biomed Online 6, 439-443.

Van Asselt KM and Kok HS. Age at menopause. A genetic-epidemiological study. Academic Thesis, Utrecht 2003.

Van Asselt KM, Kok HS, Pearson PL et al (2004a) Heritability of menopausal age in mothers and daughters. Fertil Steril 82, 1348-1351.

Van Asselt KM, Kok HS, Putter H et al (2004b) Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative trait loci influencing variation in human menopausal age. Am J Hum Genet 74, 444-453.

Van Montfrans JM, van Hooff MH, Huirne JS, Tanahatoe SJ et al(2004) Basal FSH concentrations as a marker of ovarian ageing are not related to pregnancy outcome in a general population of women over 30 years. Hum Reprod 19, 430-433.

Van Noord PA, Dubas JS, Dorland M et al (1997) Age at natural menopause in a population-based screening cohort: the role of menarche, fecundity, and lifestyle factors. Fertil Steril 68, 95-102.

Van Rooij IA, Broekmans FJ, te Velde ER, Fauser BC, Bancsi LF, de Jong FH, Themmen AP(2002) Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod 17, 3065-3071.

Van Rooij IA, Tonkelaar I, Broekmans FJ, Looman CW, Scheffer GJ, de Jong FH, Themmen AP, te Velde ER (2004). Anti-mullerian hormone is a promising predictor for the occurrence of the menopausal transition. Menopause 11, 601-6.

Van Rooij IA, Broekmans FJ, Scheffer GJ, Looman CW, Habbema JD, de Jong FH, Fauser BJ, Themmen AP, te Velde ER (2005) Serum antimullerian hormone levels best reflect the reproductive decline with age in normal women with proven fertility: A longitudinal study. Fertil Steril 83, 979-987.

van Zonneveld P, Scheffer GJ, Broekmans FJ et al (2003b) Do cycle disturbances explain the age-related decline of female fertility? Cycle characteristics of women aged over 40 years compared with a reference population of young women. Hum Reprod 18, 495-501.

van Zonneveld P, Scheffer GJ, Broekmans FJ et al (2001) Hormones and reproductive aging. Am J Human Biol 38, 83-91.

Summary

In the introduction (*chapter 1*) of this thesis static and dynamic ovarian function tests are reviewed, which supposedly can predict ovarian reserve leading to a prognosis of the reproductive potential of a woman. Ovarian reserve is currently defined as the number and quality of the follicles left at any moment in the ovary.

The aim of the studies described in this thesis was to find an answer to the following questions:

a. Which ovarian reserve test or a certain combination can predict the cohort size of small antral follicles in the early follicular phase.

b. Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of poor and hyper ovarian response in an IVF population.

c. Which ovarian reserve test or combination of ovarian reserve tests gives the best prognostic information on the probability of pregnancy in an IVF population.

We approached this questions in two ways.

1. A prospective study was conducted that compared in an integral way all currently available static ovarian reserve tests: early follicular phase blood values of follicle stimulating hormone (FSH), oestradiol (E2), inhibin B and anti-mullerian hormone (AMH), the dynamic ovarian reserve tests: the exogenous FSH ovarian reserve test (EFORT), the Clomiphene Citrate Challenge Test (CCCT), the ultrasound tests: antral follicle count (AFC), basal ovarian volume (BOV) and the intercycle variability of test results with regard to the prediction of the ovarian response after ovarian hyperstimulation in an IVF treatment. The results of this study are reported in chapters 2, 3, 4, 5 and 6.

2. A systematic review of the literature was provided including an *a priori* protocolised information retrieval on all currently available and applied tests, namely early follicular phase blood values of follicle stimulating hormone (FSH), oestradiol, inhibin B and anti-mullerian hormone (AMH), the antral follicle count (AFC), the ovarian volume and the ovarian blood flow and furthermore the clomiphene citrate challenge test (CCCT), the exogenous FSH ovarian reserve test (EFORT) and the gonadotropin releasing hormone agonist stimulation test (GAST) as measures to determine ovarian reserve and their capability to predict ovarian response and chance of pregnancy. This systematic review is reported in chapter 7.

Chapter 2 presents the comparison between the endocrine tests, Clomiphene citrate Challenge Test (CCCT), Exogenous FSH Ovarian Reserve Test (EFORT) and basal FSH, basal E2, basal Inhibin B as an integral part of all CCCT's and EFORT's, with respect to their ability to estimate the stimulable cohort of follicles in the ovaries (ovarian reserve) and analysis which test or combination of tests would give the best prediction of ovarian reserve. One hundred and ten regularly menstruating patients, aged 18-39 years, participated in this prospective

study, randomized, by a computer designed 4-blocks system study into two groups. Fifty six patients underwent a CCCT, and 54 patients underwent an EFORT. In all patients, the test was followed by an IVF treatment. The result of ovarian hyperstimulation during IVF treatment, expressed by the total number of follicles, was used as gold standard.
We showed that the best prediction of ovarian reserve was seen, when E2-increment and Inhibin B-increment were used simultaneously in a stepforward multiple regression prediction model. The CCCT could not be used in a prediction model.This findings indicates that the EFORT is the endocrine test which gives the best prediction of ovarian reserve in a linear way.

Chapter 3 reports the results of a comparison between the Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of poor and hyper responders to in vitro fertilization (IVF). We defined a 'poor' ovarian response as less than 6 oocytes after ovarian hyperstimulation in an IVF treatment and a 'hyper' response as more than 20 oocytes after such an IVF treatment. We showed that the best predictor for poor response is the CCCT. Multiple logistic regression analysis did not produce a better model in terms of improving the prediction of poor response. For hyper response, univariate logistic regression showed that the best predictor is the inhibin B-increment in the EFORT, but with a low maximal accuray of 0.78. Again, multiple logistic regression analysis did not produce a better model in terms of predicting hyper response. It means that the CCCT is superior for identification of poor responders and the EFORT (Inhibin B-increment) is superior for prediction of hyper response at cost of a high rate of false positives and neither of the two tests seem adequate to act alone for identification of both poor and hyper responders.

Chapter 4 describes the intercycle variability (ICV) of bFSH, CCCT and EFORT and secondarily the assessing of the influence of the variability of these ovarian reserve tests. Eighty five regularly menstruating patients, aged 18-39 years, participated in this prospective study, randomized, by a computer designed 4-blocks system into two groups. Fourthy three patients underwent a CCCT, and 42 patients underwent an EFORT. The bFSH level was determined as an integral part of all CCCT's and EFORT's.
Each test was performed 1-4 times in subsequent cycles, one test per cycle. During the first 3 cycles patients were treated with intra uterine inseminations (IUI). Follicle number and oocyte yield during IVF hyperstimulation in the 4th cycle were taken as measure for ovarian reserve.
We found that the intercycle variability of the Inhibin-B increment and the estradiol increment in the EFORT is stable in consecutive cycles, which indicates that this reproducible test is a more reliable tool for determination of ovarian reserve than bFSH and CCCT and that women with limited ovarian reserve show a strong intercycle variability of bFSH and FSH response to clomiphene.

Chapter 5 reports the role for the measurement of the basal ovarian volume (BOV) and antral follicle count (AFC) by transvaginal ultrasound for the prediction of ovarian reserve in comparison with the currently used endocrine ovarian reserve tests. The best prediction of ovarian reserve was seen in a multiple regression prediction model that included, AFC, Inhibin B-increment in the EFORT and BOV simultaneously.

Univariate logistic regression showed that the best predictors for poor response were the CCCT (ROC-AUC = 0.87), the bFSH (ROC-AUC = 0.83) and the AFC (ROC-AUC = 0.83). Multiple logistic regression analysis did not produce a better model in terms of improving the prediction of poor response. For hyper response, univariate logistic regression showed that the best predictors were AFC (ROC-AUC = 0.92) and the inhibin B-increment in the EFORT (ROC-AUC = 0.92), but AFC had better test characteristics, namely a sensitivity of 82 % and a specificity 89 %. Multiple logistic regression analysis did not produce a better model in terms of predicting hyper response.
In conclusion AFC performs well as a test for ovarian response being superior or at least similar to complex expensive and time consuming endocrine tests.

In *Chapter 6* we prospectively assessed the significance of serum basal Anti-Müllerian hormone (bAMH) as a novel test for ovarian reserve in an IVF population of 110 patients and compared its predictive performance with most of the established ovarian reserve tests. The outcome measures were the ovarian response after ovarian hyperstimulation in an IVF treatment expressed as the total number of stimulated follicles, retrieved oocytes and ongoing pregnancies.
We showed that the predictive value of AMH for poor response is comparable with that of AFC, but unfortunately not for the prediction of hyper responders. Included into the stepwise forward multiple regression model bAMH did not have additive value to a combination of the Inhibin B-increment in the EFORT and BOV, which led to the most optimal prediction model with regard to ovarian response. The prediction of the occurrence of pregnancy, is very limited for all tests.
But there are potential advantages of using bAMH over AFC or the CCCT, because AMH can be measured throughout the cycle in contrast to the other parameters, which can ony be determined in the early follicular phase. This study supported this phenomenon, because we did not see a change in the level of AMH after an acute endogenous rise in FSH (CCCT) and an acute exogenous rise in FSH (EFORT).
In conclusion AMH is comparable with other commonly used ovarian reserve test, but is probably most applicable in general practice, because it can be measured throughout the cycle, an advantage for both patients and clinicians. The great advantage of AFC over any other test is its potential usefulness for its ability to concomitantly predict low and high responders.

Chapter 7 presents the first comprehensive systematic literature review, including an *a priori* protocolized information retrieval on all currently available and applied tests, namely early-follicular-phase blood values of FSH, estradiol, inhibin B and anti-Müllerian hormone (AMH), the antral follicle count (AFC), the ovarian volume (OVVOL) and the ovarian blood flow, and furthermore the Clomiphene Citrate Challenge Test (CCCT), the exogenous FSH ORT (EFORT) and the gonadotrophin agonist stimulation test (GAST), all as measures to predict ovarian response and chance of pregnancy. We provided, where possible, an integrated receiver operating characteristic (ROC) analysis and curve of all individual evaluated published papers of each test, as well as a formal judgement upon the clinical value. Our analysis shows that the ORTs known to date have only modest-to-poor predictive properties and are therefore far from suitable for relevant clinical use. Accuracy of testing for the occurrence of poor ovarian response to hyperstimulation appears to be modest. Whether the a *priori* identification of

actual poor responders in the first IVF cycle has any prognostic value for their chances of conception in the course of a series of IVF cycles remains to be established. The accuracy of predicting the occurrence of pregnancy is very limited. If a high threshold is used, to prevent couples from wrongly being refused IVF, a very small minority of IVF-indicated cases (~3%) are identified as having unfavourable prospects in an IVF treatment cycle. Although mostly inexpensive and not very demanding, the use of any ORT for outcome prediction cannot be supported. As poor ovarian response will provide some information on OR status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy.

In *Chapter 8* the findings of the studies described in this thesis are discussed and the three research questions are answered The ideal ovarian reserve test for physicians should identify a substantial percentage of IVF indicated cases, who have a practically zero chance of becoming pregnant in an IVF programme. From the prospective study (*chapter 2-6*) in this thesis and the systematic reviews (*chapter7*) presented in this thesis it can be concluded that the ovarian reserve tests known to date have only minimal predictive properties and are therefore far from suitable for relevant clinical use, especially for the prediction of the occurrence of pregnancy. This implies that the use of the test as a method to deny treatment to assumed ovarian aged women should be declined. Accuracy of testing for the occurrence of poor ovarian response to hyperstimulation appeared to be clearly better than that for pregnancy. For the prediction of the cohort size of small antral follicles in the early follicular phase in a quantitative way, we could built a nice model with 3 ovarian reserve tests, but this is less usable in clinical practice. Accurate poor response prediction may be essential to identify poor responders in the first IVF cycle that suffer from severely diminished ovarian reserve. These cases, will show a very much reduced chance of pregnancy in subsequent cycles and continuation of treatment should be denied. Therefore entering the first cycle of IVF without any prior testing seems to be the preferable strategy. It should be remembered that the purpose of any ovarian reserve test is the identification of women with poor ovarian reserve for her age. This implies that chronological age always is the first step in ovarian reserve assessment.

Nederlandse samenvatting

In de introductie (h*oofdstuk 1*) van dit proefschrift worden de statische -en dynamische ovariële functie testen besproken, die de ovariële reserve kunnen voorspellen waarmee een prognose gemaakt kan worden van de reproductieve status van een vrouw. De ovariële reserve wordt gedefinieerd als het aantal follikels/eicellen en de kwaliteit van deze follikels/eicellen die in het ovarium aanwezig zijn.
Het doel van het onderzoek uit dit proefschrift was om een antwoord te vinden op de volgende vragen:

a. Welke ovariële reserve test of combinatie van deze testen geeft de beste voorspelling van het aantal eiblaasjes in de vroeg folliculaire fase van de cyclus.
b. Welke ovariële reserve test of combinatie van deze testen geeft de beste prognostische informatie over de kans op een lage, danwel hoge ovariële respons in een IVF behandeling.
c. Welke ovariële reserve test of combinatie van deze testen geeft de beste prognostische informatie over de kans op zwangerschap in een IVF behandeling.

We hebben deze vraag op twee manieren benaderd.

1. Er werd een prospectieve studie gedaan waarin alle gangbare statische ovariële reserve testen: vroeg folliculaire fase gemeten serum follikelstimulerend hormoon (FSH), oestradiol (E2), inhibine B, anti-Müllerian hormoon (AMH), dynamische ovariële reserve testen: exogene FSH ovariële reserve test (EFORT), clomifeen citraat belastings test (CCCT), vroeg folliculaire echoonderzoeken: antrale follikel meting (AFC), basale ovarium volume (BOV) en de intercyclische variabiliteit van de ovariële reserve testen werden vergeleken met betrekking tot de voorspelling van de ovariële respons na ovariële hyperstimulatie bij een IVF-behandeling. De resultaten van dit onderzoek zijn vermeld in hoofdstuk 2, 3, 4, 5 en 6.

2. Er werd een gestructureerd overzicht van de literatuur gegeven, inclusief het voorspellend vermogen met betrekking tot een lage ovariële respons en het uitblijven van zwangerschap na IVF van alle momenteel beschikbare en uitgevoerde ovariële reserve testen, namelijk vroeg folliculaire gemeten FSH, E2, inhibine B, AMH, AFC, BOV en de ovariële doorbloeding, evenals de CCCT, EFORT en GnRH agonist stimulatie test (GAST). De resultaten van dit overzicht zijn vermeld in hoofdstuk 7.

In *hoofdstuk 2* worden de clomifeen citraat belastings test (CCCT), de exogene FSH ovariële reserve test (EFORT) en basaal FSH, basaal E2, basaal inhibine B als integraal onderdeel van alle CCCT's en EFORT's met elkaar vergeleken met betrekking tot hun vermogen om de grootte van het FSH-gevoelige follikelcohort (ovariële reserve) te voorspellen, waarbij een analyse wordt gegeven welke test of combinatie van testen de beste voorspelling geeft. Het betreft een prospectieve studie waaraan 110 patiënten van 18-39 jaar met een regelmatige cyclus hebben deelgenomen, waarbij deze patiënten in twee groepen werden gerandomiseerd, 56 patiënten ondergingen een CCCT, en 54 patiënten ondergingen een EFORT. Hierna kregen alle patiënten een IVF-behandeling. De uitkomst van de ovariële hyperstimulatie tijdens de IVF-behandeling, uitgedrukt in het totaal aantal follikels, werd gebruikt als 'gouden standaard'.

We concludeerden dat de beste voorspelling van de ovariële reserve werd verkregen wanneer de E2-toename en de inhibine B-toename in de EFORT gelijktijdig werden gebruikt in een lineair model. De CCCT of andere gangbare ovariële reserve testen werden niet geselecteerd.

In *Hoofdstuk 3* worden de resultaten beschreven van de vergelijking tussen de clomifeen citraat belastings test (CCCT) versus de exogene FSH ovariële reserve test (EFORT) als test voor de identificatie van patiënten die een slechte (< 6 oocyten) danwel overgestimuleerde ovariële respons (> 20 eicellen) hebben tijdens de in vitro fertilisatie (IVF) behandeling. We hebben aangetoond dat de CCCT het beste een slechte responder kan voorspellen. Er was geen combinatie van testen die een betere voorspelling gaf dan de CCCT. Bij een te sterke respons toonde de univariate logistische regressie aan dat de inhibine B-toename in de EFORT de beste voorspelling gaf, maar met een lage maximale nauwkeurigheid van 0.78, dient deze test derhalve niet gebruikt te worden voor de voorspelling van ovariële hyper respons. Ook hier gaf een combinatie van testen geen beter model voor de voorspelling van een ovariële hyper respons.

Hoofdstuk 4 beschrijft de intercyclische variabiliteit (ICV) van bFSH, CCCT en EFORT, en wat het voorspellende vermogen van deze variabiliteit is op de ovariële reserve. Het betreft een prospectieve studie waaraan 85 patiënten van 18-39 jaar met een regelmatige menstruatie deelnamen, waarbij deze patiënten in 2 groepen weren gerandomiseerd. Elke test werd 1-4 maal gedurende achtereenvolgende cycli uitgevoerd, één test per cyclus. Gedurende de eerste 3 cycli werden de patiënten behandeld met intra-uteriene inseminatie (IUI). De 4^{e} cyclus werd gevolgd door een IVF behandeling, waarbij het aantal follikels en oöcyten die ontstonden bij de ovariële hyperstimulatie werden gebruikt als 'gouden standaard'. We concludeerden dat de intercyclische variabiliteit van de inhibine B-toename en de E2- toename in de EFORT bij achtereenvolgende cycli klein was, hetgeen betekent dat het een goede reproduceerbare test is. Daarbij concludeerden we ook dat een hoge ICV van de bFSH en CCCT correleerde met een lagere ovariële reserve, m.a.w. de ICV van de bFSH en CCCT zou kunnen worden gebruikt als ovariële reserve test.

Hoofdstuk 5 beschrijft de rol van het basaal gemeten echografische ovariële volume (BOV) en antrale follikel meting (AFC) t.a.v. de voorspelling van de ovariële reserve. Deze echografische testen werden vergeleken met de in hoofdstuk 2, 3 en 4 beschreven endocrinologische ovariële reserve testen. Na het uitvoeren van een multivariate regressie-ananlyse bleek AFC, inhibine B-toename in de EFORT en BOV in een model, de beste kwantitatieve voorspelling van de ovariële reserve te geven. Univariate logistische regressie toonde aan dat de CCCT (ROC-AUC = 0,87), bFSH (ROC-AUC = 0,83) en AFC (ROC-AUC = 0,83) het beste een lage ovariële respons kunnen voorspellen. Multivariate logistische regressieanalyse leverde geen beter model op voor de voorspelling van een lage respons. Voor de voorspelling van een hyper respons waren AFC (ROC-AUC = 0,92) en inhibine B-toename in de EFORT (ROC-AUC = 0,92) de beste voorspellers, maar AFC had betere testeigenschappen, een sensitiviteit van 82% en een specificiteit van 89%. Multiple logistische regressieanalyse leverde geen beter model op voor de voorspelling van een hyper respons. Hieruit kan geconcludeerd worden dat AFC een goede test is voor de ovariële reserve en beter is dan, of tenminste net zo goed is als ingewikkelde, dure en tijdrovende endocrinologische testen.

Hoofdstuk 6 beschrijft een prospectieve studie, waarbij de rol van basaal anti-Müllerian hormoon (bAMH) als een nieuwe marker voor ovariële reserve in een IVF-populatie van 110 patiënten wordt beschreven. AMH werd vergeleken met de in hoofdstuk 2, 3, 4 en 5 beschreven ovariële reserve testen. De eindpunten van deze studie zijn de ovariële respons na hyperstimulatie van de ovaria in een IVF-behandeling en een doorgaande zwangerschap.
De voorspellende waarde van AMH voor een lage ovariële respons is vergelijkbaar met AFC, helaas geeft AMH geen goede voorspelling op een hyper respons. Alle testen konden een eventuele zwangerschap maar in beperkte mate voorspellen. Een groot voordeel van AMH is dat het op ieder moment van de cyclus bepaald kan worden, hetgeen een praktisch voordeel is voor arts en patiënt. Dit wordt door deze studie ondersteund, daar de AMH concentraties niet werden beïnvloed door een plotseling stijging van FSH en LH spiegels na het toedienen van een hoge dosis recFSH of clomifeen citraat. Het grote voordeel van AFC is dat het een goede voorspeller is voor zowel een lage ovariële respons als een hyper ovariële respons.

Hoofdstuk 7 beschrijft een gestructureerd overzicht van de literatuur, inclusief het voorspellend vermogen met betrekking tot een lage ovariële respons en het uitblijven van zwangerschap na IVF van alle momenteel beschikbare en uitgevoerde ovariële reserve testen, namelijk vroeg folliculair gemeten FSH, E2, inhibine B, AMH, AFC, BOV en de ovariële doorbloeding, evenals de CCCT, EFORT en GAST. De resultaten van dit overzicht zijn vermeld in hoofdstuk 7. Van alle studies die geselecteerd werden, werden geïntegreerde ROC-curves gemaakt. Vanwege de heterogeniciteit tussen de diverse studies was het berekenen van een puntschatter voor sensitiviteit en specificiteit niet altijd mogelijk. Uit de analyses blijkt dat de huidige ovariële reserve testen slechts een zeer bescheiden vermogen hebben en zijn daarom niet klinisch bruikbaar. Dit geldt met name voor het voorspellend vermogen ten aanzien van het uitblijven van een zwangerschap. Het voorspellend vermogen van deze testen met betrekking tot een lage respons is veel beter.

In *hoofdstuk 8* worden de bevindingen besproken van de studies die zijn beschreven in dit proefschrift en de 3 onderzoeksvragen worden beantwoord. De ideale ovariële reserve test moet een aanzienlijk deel van die IVF patiënten identificeren, die nagenoeg geen kans meer hebben zwanger te worden met behulp van IVF. Uit de prospectieve studies (*hoofdstuk 2-6*) en het systematische overzicht en meta-analyse van de ovariële reserve testen (*hoofdstuk 7*) in dit proefschrift, kan geconcludeerd worden dat de huidige ovariële reserve testen slechts een zeer bescheiden voorspellend vermogen hebben en zijn daarom klinisch niet bruikbaar, met name niet t.a.v. de voorspelling van zwangerschap. Het gebruik van deze testen met als doel patiënten een behandeling te onthouden moet dan ook vermeden worden. Het voorspellend vermogen van deze testen met betrekking tot lage respons is duidelijk beter. De kwantitatieve voorspelling van de ovariële reserve middels een lineair model is goed, echter deze is niet goed te gebruiken in de dagelijkse praktijk. Het voorspellen van lage respons kan belangrijk zijn om die lage responders in de eerste behandelingscyclus te identificeren, die daadwerkelijk een verminderde ovariële reserve hebben. Deze patiënten hebben een duidelijk verminderde kans op zwangerschap in de vervolgcycli. Het voortzetten van de behandeling moet achterwege gelaten worden. Ons inziens moet een eerste IVF cyclus, zonder voorafgaande ovariële reserve testen, de nieuwe behandel-strategie worden, waarin leeftijd eveneens een belangrijke rol speelt.

Dankwoord

Het krijgen van een indonesische lunch, een uitnodiging op een marokkaans feest en nog vele leuke ander attenties is toch niet wat je verwacht als je een klinisch onderzoek gaat doen. Maar het zijn juist deze dingen waar je door wordt gedreven. Ik wil alle vrouwen die hun enthousiaste medewerking hebben verleend hier dan ook hartelijk voor bedanken.
Daarnaast is dit proefschrift zonder de enthousiaste inzet van vele anderen niet mogelijk geweest. De volgende personen wil ik met name noemen.

Prof. Dr. J. Schoemaker, beste Joop, trots ben ik dat het eindelijk af is. Helaas is het me niet gelukt om binnen jouw tijd te promoveren. Toen mij dit duidelijk werd, moest ik wel even een paar keer slikken, maar gelukkig heeft Nils jouw taak bijzonder goed overgenomen. Dank voor je inwijding in de endocrinologie, in het onderzoek, je steun en zelfs was je goed in het drogen van gefrustreerde tranen. Ook Toke heel erg bedankt dat ik jullie zo vaak heb mogen storen in jullie paradijsje in Drenthe.

Prof. Dr. R. Homburg, dear Roy, last but not least!!!!!

Dr. C.B. Lambalk, beste Nils, DANK, DANK, DANK dat je me geadopteerd hebt. Je hebt me door je mateloze enthousiasme erdoorheen gesleept, zonder jou was het me nooit gelukt. Vele ideeën hebben we gedeeld en daar zijn er gelukkig veel van opgeschreven, maar gelukkig ook een hele hoop niet. Ik zal je missen!

Dr. R. Schats, beste Roel, jij bent de hoeksteen van het onderzoek geweest en hebt me de mogelijkheid geboden om dit onderzoek op jouw IVF centrum te laten doen plaatsvinden. Hiervoor heel veel dank.

J. McDonnell, beste Joseph, als statisticus was je het niet altijd met me eens, maar kwamen we er altijd weer uit. Gelukkig heeft jouw nuchtere Australische blik een zogende moeder met kind achter de computer niet afgeschrikt. Ik heb veel van je geleerd.

De leden van de promotie commissie, Dr. F. Broekmans, Prof. dr. H.A.M. Brölmann, Prof. dr. M.J. Heineman, Prof. dr. F. de Jong, Dr. G.J.E. Oosterhuis, Dr. R. Schats, Prof. dr. F. Scheele, bedank ik voor het kritisch en snel doornemen van dit proefschrift.

Judith Huirne, lieve Judith, samen zijn we aan deze lange weg van diepe dalen en hoge bergen begonnen en samen maken we het nu ook af. Jammer dat onze assistenten tijd er bijna opzit, maar onze vriendschap zal nooit verloren gaan. Ik vind het fantastisch dat we nu samen promoveren.

Sandra Tanahatou, lieve Sandra, zonder echte maatjes kom je er niet en jij bent er een van. Je bent een echte vriendin geworden. Ik hoop dat onze ‘maatschap’ dan ook nooit verloren gaat. Ik ben vereerd dat je mijn paranimf wilt zijn.

Hans van Weering, lieve vriend, je wilde niet op mijn maiden party komen, maar gelukkig wil je wel mijn paranimf zijn en het colpobusje gaat echt nog wel een keer rijden!

Mariet Elting, lieve Mariet, jij was mijn onderzoeksmaatje en we hebben dan ook een super tijd gehad op het IVF. Ben blij dat we nog steeds contact hebben.

Ruth Janssens, lieve Ruth, onze band is for ever.

Madeleine Evers, beste Madeleine, dank voor alle ondersteuning, met name van de laatste loodjes die wel héééééél zwaar waren.

IVF: Renee Dirks en Julie Knieriem, biologen, verpleegkundigen, secretariaat, en labmedewerkers, wat hebben we ongelofelijk veel gelachen. Nog steeds krijg ik een soort ik-ben-weer-thuis gevoel als ik op het IVF centrum ben. Iedereen heel erg bedankt voor de onwijs leuke tijd die ik gehad heb en natuurlijk ook veel dank voor het opvangen en recruteren van mijn patiënten.

Lieve vriendinnen, dank voor jullie oprechte belangstelling. Het is gelukkig af, tijd voor een feestje!!!!!!

Anneke Kwee, lieve zus en vriendin, ben erg blij met een zus zoals jij, kon je altijd bellen voor advies en je hebt altijd meegeleefd. Maar nu is er echt tijd om leuke dingen te gaan doen (zie p. 194 Caesarean section in the Netherlands, A. Kwee, oktober 2005).

Lieve daddy en mamma, zonder jullie onvoorwaardelijke steun en grenzeloze liefde was dit boekje er niet gekomen. Dat jullie iedere dinsdag weer kwamen tijdens mijn ouderschaps verlofdag om op te passen, zodat ik achter de computer kon duiken of naar de VU kon gaan, heeft dit mogelijk gemaakt.

Lieve Puck, Keet en Ties, bij jullie valt echt alles in het niet, daarbij komt ook nog dat zeker zonder jullie komst dit proefschrift er nooit geweest was. Elk zwangerschapsverlof was goed voor gemiddeld 2 artikelen.

Lieve Paul, woorden schieten te kort!!!!! Wat had ik zonder jou gemoeten.

List of publications

Schats R, Vink J, Kwee J. IUI, IVF, ICSI de grens bij 40 jaar. *Ned Tijdschr Obstet Gynaecol 1999;112:52-6.*

Elting MW, Kwee J, Schats R, Rekers-Mombarg LT, Schoemaker J. The rise of estradiol and inhibin B after acute stimulation with follicle-stimulating hormone predict the follicle cohort size in women with polycystic ovary syndrome, regularly menstruating women with polycystic ovaries, and regularly menstruating women with normal ovaries. *J Clin Endocrinol metab 2001;86:1589-95.*

Elting MW, Kwee J, Korsen TJ, Rekers-Mombarg LT, Schoemaker J. Aging women with polycystic ovary syndrome who achieve regular menstrual cycles have a smaller follicle cohort than those who continue to have irregular cycles. *Fertil Steril 2003;79:1154-60.*

Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB, Schoemaker J. Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment : results of a prospective randomized study. *Human Reproduction 2003;18:1422-7.*

Kwee J, Schats R, McDonnell J, Lambalk CB, Schoemaker J. Intercycle variability of ovarian reserve tests: results of a prospective randomized study. *Human Reproduction 2004;19:590-5.*

Reply: Kwee J and Lambalk CB. Variability of ovarian reserve tests *Human Reproduction* 2004;19*: 2171.*

Kwee J, Schats R, McDonnell J, Schoemaker J, Lambalk CB. The Clomiphene Citrate Challenge Test (CCCT) versus the Exogenous Follicle stimulation hormone Ovarian Reserve Test (EFORT) as single test for identification of low and hyperresponders to in vitro fertilization (IVF). *Fertil Steril 2006; 85:1714-22.*

Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. *Human Reproduction Update 2006;6:685-718.*

Curriculum Vitae

Curriculum vitae

23-07-1969	Geboren te Winschoten
1987-1987	VWO, Winschoter Scholen gemeenschap
1987- 1993	Doctoraalstudie geneeskunde, Rijksuniversiteit Groningen
1996-1996	Coschappen en Artsexamen, Vrije Universiteit Amsterdam
2000-2000	AIO gynaecologie, IVF centrum VUMC
2000- heden	opleiding gynaecologie VUMC, Amsterdam en St Lucas Andreas Ziekenhuis, Amsterdam (opleiders: Prof. Dr. H.P. van Geijn, Prof. Dr. H.A.M. Brölmann, Dr. J.Th.M. van der Schoot, Prof. Dr. F. Scheele)

Janet is getrouwd met Paul Holtrop en samen hebben ze een dochter Puck (18-12-2001), een dochter Keet (20-10-2003) en een zoon Ties (11-12-2005).